DOI: 10.1534/genetics.104.026500

Note

The HKA Test Revisited: A Maximum-Likelihood-Ratio Test of the Standard Neutral Model

Stephen I. Wright[1] and Brian Charlesworth

Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Ashworth Laboratories, Edinburgh EH9 3JT, United Kingdom

Manuscript received January 15, 2004
Accepted for publication June 25, 2004

ABSTRACT

We present a maximum-likelihood-ratio test of the standard neutral model, using multilocus data on polymorphism within species and divergence between species. The model is based on the Hudson-Kreitman-Aguadé (HKA) test, but allows for an explicit test of selection at individual loci in a multilocus framework. We use coalescent simulations to show that the likelihood-ratio test statistic is conservative, particularly when the assumption of no recombination is violated. Application of the method to polymorphism data from 18 loci from a population of *Arabidopsis lyrata* provides significant evidence for a balanced polymorphism at a candidate locus thought to be linked to the centromere. The method is also applied to polymorphism data in maize, providing support for the hypothesis of directional selection on genes in the starch pathway.

THE neutral theory of molecular evolution predicts that the amount of within-species diversity should be correlated with levels of between-species divergence, due to the dependence of both on the neutral mutation rate (Kimura 1983). The widely used HKA test (Hudson *et al.* 1987) evaluates the fit of polymorphism and divergence data to this prediction, as a test for natural selection against the null hypothesis of neutrality. The method involves the use of within-species polymorphism data on a sample from one species and sequence divergence from a related species, so that the relative amounts of polymorphism and divergence can be compared across loci. When applied to multilocus data, the HKA test assesses the overall fit of the data to a neutral model that assumes the same ratios of polymorphism and divergence at each locus. The marginal contributions of each locus to the multilocus chi-square statistic, or the results of multiple pairwise HKA tests, are often used to assess which loci contribute most to any observed departure from neutrality (*e.g.*, Moore and Purugganan 2003). This approach, however, does not provide a way of rigorously comparing different models, for example, to test for selection at a specific locus.

Here, we develop a maximum-likelihood method for analyzing polymorphism and divergence, using the HKA framework. As with the HKA test, the method assumes no recombination within loci but free recombination between loci, and it assumes that the ancestral population size of the species from which polymorphism data were obtained was the same as the current population size. The method is based on the assumption that loci are statistically independent. The likelihood (Edwards 1972) of the observed numbers of segregating sites (S_i) and pairwise divergence (D_i) across each locus i, for a total of r loci sampled from a population, is then given by

$$L = \prod_{i=1}^{r} L_i(\theta_i|S_i)\, L_i(\theta_i,\ T|D_i), \qquad (1)$$

where $\theta_i = 4N_e u_i$, N_e is the effective population size, u_i is the locus-specific mutation rate, and T is the divergence time of the two species, in units of $2N_e$ generations.

As with the HKA test, the method also assumes that polymorphism and divergence are independent. Although a slight nonindependence is expected between polymorphism and divergence for a given locus, the effect is very weak, and violations of the independence assumption are expected to lead to slightly conservative tests (Hudson *et al.* 1987). Under the assumption of no recombination and the standard neutral model of coalescence in a panmictic population, the likelihood of θ_i given S_i and sample size n is proportional to the probability of S_i given θ_i and sample size $n(P_n(S_i|\theta_i))$ according to the recursion

[1]*Corresponding author:* Department of Biology, York University, 4700 Keele St., Toronto, ON M3J 1P3, Canada.
E-mail: stephenw@yorku.ca

TABLE 1

Maximum-likelihood analysis of synonymous polymorphism and divergence for *A. lyrata*

Model	Description	k^a	ln L	Model comparison	Likelihood-ratio statistic (d.f.)	P^b
1	Fixed mutation, no selection	—	−131.8		—	
2	Free mutation, no selection	—	−95.9	M2 *vs.* M1	71.8 (17)	8.4×10^{-9}
3	Fixed mutation, selection	9.1	−119.8	M3 *vs.* M1	24 (1)	9.6×10^{-7}
4	Free mutation, selection	4.3	−91.4	M4 *vs.* M3	56.8 (17)	3.5×10^{-6}
				M4 *vs.* M2	9 (1)	2.7×10^{-3}

[a] Selection parameter for the gene AT1G36310.
[b] Probability of likelihood-ratio test, assuming the χ^2 distribution.

$$L_n(\theta_i|S) \propto P_n(S_i|\theta_i) = \sum_{j=0}^{S_i} P_{n-1}(S_i - j|\theta_i)\, Q_n(j|\theta_i) \quad (2)$$

(Hudson 1990), where

$$P_2(S|\theta) = \left(\frac{\theta}{1+\theta}\right)^S \left(\frac{1}{1+\theta}\right)$$

and

$$Q_n(S|\theta) = \left(\frac{\theta}{\theta + n - 1}\right)^S \left(\frac{n-1}{\theta + n - 1}\right).$$

The likelihood of the number of pairwise differences between a random sequence chosen from each species is given by Takahata *et al.* (1995):

$$L(T, \theta_i|D_i) \propto P(D|\theta_i, T) = \left\{1 - \left(\frac{\theta_i}{1+\theta_i}\right)\right\} \left(\frac{\theta_i}{1+\theta_i}\right)^{D_i} e^{-\theta_i T} \sum_{m=0}^{D_i} \frac{1}{m!} \left(\frac{(1+\theta_i)\theta_i T}{\theta_i}\right)^m. \quad (3)$$

The full neutral model thus has $r + 1$ parameters: a θ parameter for each locus, plus the shared divergence time parameter. The maximum log-likelihood under this model can thus be compared with the maximum log-likelihood with a common u parameter for all loci ($u_1 = u_2 = \ldots = u_r$) to provide a likelihood-ratio test for the null hypothesis of a common mutation rate for all loci.

Within the HKA framework, selection acts by uncoupling polymorphism from divergence; positive selection will reduce the θ parameter for polymorphism relative to that for divergence, since the fixation of a favorable allele will reduce levels of diversity at linked sites by hitchhiking, while long-term balancing selection will have the opposite effect of maintaining elevated levels of diversity at linked neutral sites (Kreitman 2000). A model with a single selected locus (the rth locus) thus has one additional scaling parameter k:

$$L = L(k\theta_r|S_r)\, L(\theta_r, T|D_r) \prod_{i=1}^{r-1} L(\theta_i|S_i)\, L(\theta_i, T|D_i). \quad (4)$$

In this model, k measures the degree to which diversity is increased or decreased by the action of selection at locus r. The fit under this model can then be compared to that under the alternative of no selection at the rth locus by a likelihood-ratio test.

To maximize the likelihood of the models, we use a Monte Carlo Markov chain following an approach of simulated tempering similar to that of McVean and Vieira (2001), making use of the Metropolis-Hastings algorithm (Metropolis *et al.* 1953; Hastings 1970). Briefly, a parameter is chosen at random, and the value of this parameter is incremented using a uniform distribution between −λ and +λ, where λ is a predefined increment. The likelihood of the data under the new parameter combination is then calculated, and the change is accepted if the likelihood of the new parameter set is greater. If the likelihood is less than the previous likelihood, the change is accepted with probability proportional to the difference in log-likelihoods:

$$P[\text{accept}] = \exp(f(\ln L_{\text{new}} - \ln L_{\text{old}})). \quad (5)$$

For changes with lower likelihood, the probability of acceptance is determined by multiplying the difference in log-likelihood by a factor f of 50. After 10,000 iterations, this acceptance probability is reduced to a factor f of 0.5 times the difference in log-likelihoods for 1000 iterations and then returned to the original value. The Markov chain is run multiple times, using different starting parameters and different random-number seeds to ensure convergence. To test a difference between two models that yield maximum likelihoods L_1 and L_2, we use the likelihood-ratio test statistic $2(\ln L_2 - \ln L_1)$, *i.e.*, twice the difference in log-likelihoods between two models. A program written in C++ to carry out the method is available for download at www.yorku.ca/stephenw.

Here we consider four likelihood models:

Model 1 (1 free parameter): fixed mutation ($u_1 = u_2 = \ldots u_r$), no selection

$$(k_1 = k_2 = \ldots k_r = 1)$$

Model 2 (r free parameters, where r = number of loci): free mutation (all θ_i estimated independently), no selection ($k_1 = k_2 = \ldots . k_r = 1$)

TABLE 2

Results of simulations of parameter estimates using the maximum-likelihood estimators based on polymorphism and divergence

		No intragenic recombination			Intragenic recombination		
Parameter	Value[a]	Mean	Median	Standard error	Mean	Median	Standard error
T	17.83	18.50	18.18	0.0982	18.22	18.04	0.0840
θ_1	0.0084	0.0084	0.0082	7.4×10^{-5}	0.0083	0.0081	6.6×10^{-5}
θ_2	0.0070	0.0068	0.0066	6.4×10^{-5}	0.0069	0.0069	6.0×10^{-5}
θ_3	0.0067	0.0066	0.0066	6.5×10^{-5}	0.0066	0.0065	6.1×10^{-5}
θ_4	0.0066	0.0066	0.0064	6.2×10^{-5}	0.0066	0.0065	5.7×10^{-5}
θ_5	0.0049	0.0049	0.0048	4.9×10^{-5}	0.0049	0.0049	4.7×10^{-5}
θ_6	0.0085	0.0085	0.0084	7.0×10^{-5}	0.0085	0.0083	6.5×10^{-5}
θ_7	0.0052	0.0052	0.0050	5.4×10^{-5}	0.0052	0.0050	5.0×10^{-5}
θ_8	0.0068	0.0068	0.0067	6.0×10^{-5}	0.0069	0.0068	5.2×10^{-5}
θ_9	0.0090	0.0089	0.0087	7.5×10^{-5}	0.0090	0.0087	7.0×10^{-5}
θ_{10}	0.0282	0.0280	0.0275	1.7×10^{-4}	0.0284	0.0278	1.6×10^{-4}
θ_{11}	0.0105	0.0103	0.0101	8.1×10^{-5}	0.0106	0.0104	8.1×10^{-5}
θ_{12}	0.0127	0.0126	0.0124	9.2×10^{-5}	0.0129	0.0127	9.3×10^{-5}
θ_{13}	0.0079	0.0078	0.0075	6.4×10^{-5}	0.0079	0.0078	6.2×10^{-5}
θ_{14}	0.0077	0.0076	0.0073	6.7×10^{-5}	0.0076	0.0076	6.3×10^{-5}
θ_{15}	0.0131	0.0131	0.0129	9.1×10^{-5}	0.0131	0.0130	9.0×10^{-5}
θ_{16}	0.0115	0.0113	0.0112	8.8×10^{-5}	0.0115	0.0114	8.0×10^{-5}
θ_{17}	0.0087	0.0086	0.0083	7.8×10^{-5}	0.0087	0.0085	7.4×10^{-5}
θ_{18}	0.0054	0.0054	0.0052	5.2×10^{-5}	0.0054	0.0053	5.1×10^{-5}

[a] Values of divergence time and population mutation parameters used in simulations.

Model 3 (1 free parameter): fixed mutation ($u_1 = u_2 = \ldots u_r$), selection at candidate locus l (k_l estimated, $k_2 = k_3 = \ldots . k_{i \neq 1} = 1$)

Model 4 ($r + 1$ free parameters): free mutation (all u_i estimated independently), selection at candidate locus l (k_l estimated, $k_1 = k_2 = \ldots . k_{i \neq 1} = 1$).

We have applied the method to polymorphism data for 18 genes (supplementary Table 1 at http://www.genetics.org/supplemental/) from a single Icelandic population of *Arabidopsis lyrata*, using divergence estimates from *A. thaliana* (S. Wright and D. Charlesworth, unpublished data). These loci have an average of 126 synonymous sites and 3.9 segregating synonymous sites, and the sample size averaged 16.7 haploid genomes. For this data set, 100,000 chains were found to be sufficient for convergence, which took ~5 min to run on a 2.8-GHz Pentium computer. To evaluate the behavior of the model, and to assess the fit of the likelihood ratio statistic to the χ^2 approximation, we ran neutral coalescent simulations (Hudson 2002) for 18 loci, with divergence time, sample size, and θ parameter values chosen to match estimates from the *A. lyrata* data set. All coalescent simulations included a population split and a single sample from the outgroup species, so that polymorphism and divergence were both estimated from the same coalescent process. We ran two sets of 1000 simulations; the first had no intragenic recombination, and the second included intragenic recombination, with values of the population recombination parameter $\rho = 4N_e r$ estimated for each locus using genetic mapping data from *A. thaliana*, as previously described (Wright *et al.* 2003). Across loci, the assumed per-locus values of ρ ranged from 0.3 to 60, and the ratio ρ/θ ranged from 0.02 to 10. Our candidate locus for selection ("locus 10") is a gene thought to be linked to the centromere, AT1G36310, which had been hypothesized (S. Wright and D. Charlesworth, unpublished data) to be linked to a site under selection, due to increased hitchhiking in a region of reduced recombination.

To test the assumption of heterogeneity in mutation rates with this data set, we compare a fixed-rate model (model 1) with a free-rate model (model 2). Significance is assessed using a likelihood-ratio test, using χ^2 with $r - 1 = 17$ d.f. Model 2 provides a highly significant improvement to the likelihood compared with model 1, indicating the existence of heterogeneity in mutation rates across loci (Table 1). This highlights the importance of incorporating mutation rate variation into parameter estimation (*e.g.*, Wall 2003) and tests of natural selection. Simulations show that likelihood estimation of divergence and mutation parameters perform well under the neutral model, even in the presence of intragenic recombination (Table 2).

To test for selection, we first use the standard multilocus HKA test for comparison, using the program of J. Hey (http://lifesci.rutgers.edu/heylab/DistributedProgramsandData.htm#HKA). The standard multilocus HKA test for these 18 loci shows a significant departure from neutral expectation ($\chi^2 = 43.6$, $P < 0.001$), sug-

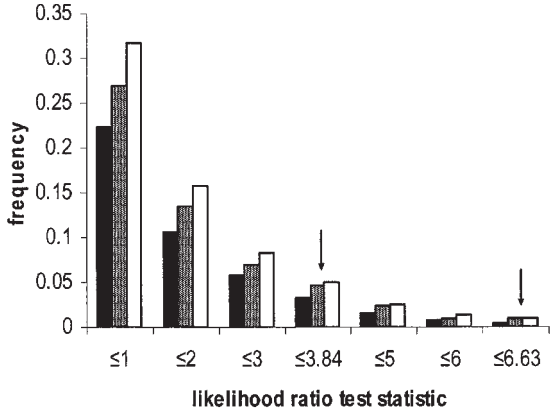

FIGURE 1.—Neutral simulation results studying the behavior of the likelihood-ratio test for selection in the presence and absence of intragenic recombination. Shown is the frequency distribution for the proportion of neutral simulations showing twice the difference in log-likelihoods calculated for the selection model (model 4 in Table 1), allowing selection on locus 10, *vs.* the neutral model (model 2 in Table 1), assuming free mutation rates across loci. Solid bars are from simulations with intragenic recombination, and shaded bars are with no intragenic recombination. Open bars are the values expected from the χ^2 distribution. Arrows show 5% and 1% significance levels.

gesting the action of natural selection at some loci. However, no individual locus shows evidence for selection using the HKA test; the maximum marginal χ^2 deviation is contributed by the polymorphism cell for the candidate locus (locus 10), which shows elevated levels of polymorphism, but this value is not significantly greater than expected under neutral coalescent simulations ($P = 0.13$). Furthermore, pairwise HKA tests comparing locus 10 to all other loci give 7 of 17 significant tests at the 5% level. While the above results are suggestive of a departure from neutrality, it is difficult to make a clear inference of selection at locus 10 using standard methods.

If we treat this centromeric locus as a candidate gene for selection using the likelihood method, comparison of a free-mutation model (model 4), which allows for selection at this locus, with the corresponding strict neutral model (model 2) gives a likelihood-ratio statistic of 9.0, which is significant under the χ^2 approximation with 1 d.f. (Table 1). Similarly, there is a highly significant improvement to the fixed-mutation-rate model when we allow for selection at the candidate locus (comparison of models 1 and 3 in Table 1). Furthermore, the maximum-likelihood estimate of the selection parameter k is 4.3, suggesting a fourfold elevation of diversity over neutral expectation at this locus.

Simulations show that the neutral model with no recombination follows the χ^2 approximation well and that the inclusion of reasonable levels of intragenic recombination makes the test more conservative (Figure 1). Under neutral coalescent simulations with intragenic recombination, the mean and median value of k are close to 1 (mean $k = 1.05$, median = 0.97), as expected. Furthermore, none of the 1000 simulations shows an estimate of k greater than that observed in this data set (Figure 2). Thus, the likelihood method shows significant evidence for a balanced polymorphism at this locus. These results suggest that the likelihood method has higher power than a multilocus HKA test to test for selection at a candidate gene.

To assess the power of the method to detect directional selection, we applied our method to a multilocus data set in maize (WHITT *et al.* 2002), which compares patterns of diversity for six genes involved in the starch pathway, which have been hypothesized to be under directional selection during domestication, with 11 neutral reference loci. Because of the considerable increase in number of sites per locus (average across loci, 795), the number of segregating sites per locus (average across loci, 26.9), and the divergence (average number of differences between species, 46.5), this analysis took considerably longer to converge (1–10 million chains) and to run (~17–100 hr on a 2.6-GHz Pentium III com-

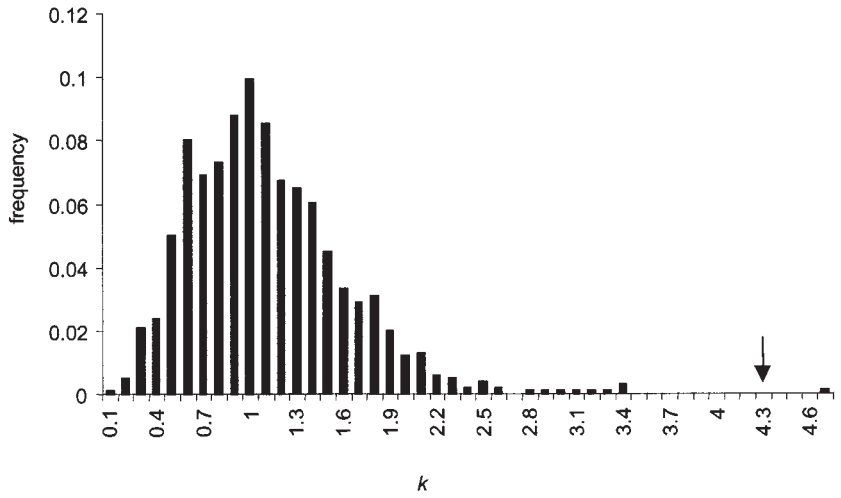

FIGURE 2.—Distribution of the maximum-likelihood estimate of the selection parameter k for locus 10, from 1000 neutral simulations with intragenic recombination. The arrow shows the value estimated for the locus AT1G36310 in the *A. lyrata* data set.

TABLE 3

Maximum-likelihood analysis of silent polymorphism in the maize data set of WHITT *et al.* (2002)

Model	Description	ln L	Comparison	Likelihood-ratio statistic (d.f.)	P	k					
						bt2	*ae1*	*su1*	*sh1*	*sh2*	*wx*
A	Neutral (all $k = 1$)	−131.4				1	1	1	1	1	1
B	Selection, 3 loci in starch pathway	−110.7	A *vs.* B	32.8 (3)	3.6×10^{-7}	1	0.054	1	0.31	0.26	1
C	Selection, 6 loci in starch pathway	−110.5	A *vs.* C	32.54 (6)	1.0×10^{-5}	2.0	0.04	0.93	0.46	0.24	1.87
			B *vs.* C	0.4 (3)	0.47						

puter) in comparison with the *A. lyrata* data set. We first compare a neutral model, where all 17 loci have $k = 1$, with a selection model, allowing all six starch pathway genes to be under selection. The likelihood-ratio test is highly significant for this comparison, showing strong evidence for selection on starch pathway genes (Table 3). However, three of the genes have maximum-likelihood estimates of k that are either close to or greater than one, suggesting that only a subset of genes may have been under directional selection. A model that allows only the 3 loci with $k < 1$ to be under selection is also highly significant in comparison with the neutral model, and the model with 6 selected loci shows no significant improvement to the likelihood in comparison with the three-gene model (Table 3).

As shown using the maize data set, the method can also easily be applied to test for selection on multiple loci, and the program allows for such models to be tested. However, the power of the method relies on the assumption that at least one locus in the data set follows the strict neutral assumptions and therefore still requires some *a priori* selection of candidate genes. With the inclusion of a significant number of neutral loci, the method should be much less sensitive to unusual loci and should incorporate the inherent uncertainty in divergence time from individual loci. This should give the likelihood method much more consistency than using a large number of pairwise tests, which have been shown to be difficult to interpret (MORIYAMA and POWELL 1996).

Although the method appears to be fairly robust to the assumption of no recombination, the power to detect a significant reduction or increase in diversity in a given region is likely to depend on the local rate of recombination and the size of the region analyzed; a signature of natural selection using this method may be unlikely in large sampled regions of high recombination, unless selection is very recent and strong. For these types of data, tests that explicitly analyze variation in patterns of diversity across a region (*e.g.*, KIM and STEPHAN 2002) would be more appropriate.

In addition to the effects of natural selection, violations from the infinite-sites model, intralocus variation in mutation rates, population subdivision, and recent changes in population size (*i.e.*, population bottlenecks and expansion) could lead to a significant likelihood-ratio test, although the extent to which the variance across loci in the ratio of diversity to divergence would be affected by these violations remains unclear. The problems associated with violations of the assumptions of the standard neutral model are shared with most other tests of selection based on polymorphism data, although in principle the multilocus comparisons associated with the HKA framework should be more robust to such violations than tests based on comparing different aspects of diversity at a single locus (*e.g.*, TAJIMA 1989; FAY and WU 2000). One important approach to addressing this question would be to first fit the multilocus data to a demographic or mutation model and then examine the distribution of the likelihood-ratio test statistic under this model. If the fit to the chi-square statistic is found to be poor, then more computationally intensive likelihoods could be estimated using simulated data rather than the given equations above, using the same basic likelihood framework. However, the method may be conservative under some demographic models; for example, population expansion is expected to reduce the variance in diversity across loci, and in this case direct use of the above method may be preferable to avoid the computational requirements of exploring by simulation a vast array of possible demographic models and mutation parameters to reach the maximum likelihoods.

We thank D. Charlesworth for helpful discussion, E. Buckler for providing data, and B. Gaut, P. Andolfatto, and two anonymous reviewers for comments on the manuscript. This research was supported by a Royal Society Professorship to B.C. and a Commonwealth Scholarship to S.W.

LITERATURE CITED

EDWARDS, A. W. F., 1972 *Likelihood.* Cambridge University Press, Cambridge, UK.

FAY, J. C., and C.-I WU, 2000 Hitchhiking under positive Darwinian selection. Genetics **155:** 1405–1413.

Hastings, W. K., 1970 Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57:** 97–109.

Hudson, R. R., 1990 Gene genealogies and the coalescent process, pp. 1–44 in *Oxford Surveys in Evolutionary Biology*, edited by J. A. D. Futuyma. Oxford University Press, New York.

Hudson, R. R., 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics **18:** 337–338.

Hudson, R. R., M. Kreitman and M. Aguadé, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153–159.

Kim, Y., and W. Stephan, 2002 Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics **160:** 765–777.

Kimura, M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, UK.

Kreitman, M., 2000 Methods to detect selection in populations with applications to the human. Annu. Rev. Genomics Hum. Genet. **1:** 539–559.

McVean, G. A., and J. Vieira, 2001 Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in Drosophila. Genetics **157:** 245–257.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, 1953 Equations of state calculations by fast computing machines. J. Chem. Phys. **21:** 1087–1092.

Moore, R. C., and M. D. Purugganan, 2003 The early stages of gene duplication. Proc. Natl. Acad. Sci. USA **100:** 15682–15687.

Moriyama, E. N., and J. R. Powell, 1996 Intraspecific nuclear DNA variation in Drosophila. Mol. Biol. Evol. **13:** 261–277.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.

Takahata, N., Y. Satta and J. Klein, 1995 Divergence time and population size in the lineage leading to modern humans. Theor. Popul. Biol. **48:** 198–221.

Wall, J. D., 2003 Estimating ancestral population sizes and divergence times. Genetics **163:** 395–404.

Whitt, S. R., L. M. Wilson, M. I. Tenaillon, B. S. Gaut and E. S. t. Buckler, 2002 Genetic diversity and selection in the maize starch pathway. Proc. Natl. Acad. Sci. USA **99:** 12959–12962.

Wright, S. I., B. Lauga and D. Charlesworth, 2003 Subdivision and haplotype structure in natural populations of Arabidopsis lyrata. Mol. Ecol. **12:** 1247–1263.

Communicating editor: A. H. Paterson