Journal Compilation. © 2008 Cambridge University Press. *Genet. Res., Camb.* (2007), **89**, *pp.* 427–432.
doi:10.1017/S0016672308009610 Printed in the United Kingdom

Genet. Res., Camb. (1987) **50**, *pp.* 245–250

Estimating the recombination parameter of a finite population model without selection

RICHARD R. HUDSON
N.I.E.H.S., P.O. Box 12233, Research Triangle Park, N.C. 27709, U.S.A.

(Received 15 April 1987 and in revised form 29 June 1987)

Summary

An estimator is proposed for the parameter C = 4Nc. where N is the population size and *c* is the recombination rate. The estimator is appropriate for use with sequence or restriction site data from random samples from within populations. Properties of the estimator are investigated for an infinite-sites neutral model using Monte Carlo simulations. The median and mode of the distribution of the estimator are close to the true value for all parameter values examined, but large data sets are required to obtain reliable estimates.

1. Introduction

In the neutral theory of molecular evolution one parameter plays a dominant role when considering within-population molecular variation. That parameter is $\theta = 4Nu$, where N is the population size and u is the neutral mutation rate. Methods of estimating this parameter and statistical properties of the estimates are well known under the neutral model (Ewens, 1979; Tajima, 1983). With intermediate levels of recombination (as in most nuclear genes) there is another parameter which plays a prominent role in this model, namely, $C = 4Nc$, where c is the recombination rate. This parameter affects, among other things, the distribution of linkage disequilibrium between sites and the variance of the number of segregating sites in samples. Unfortunately, relatively little is known about estimating C. A number of authors have considered the related problem of estimating N when c is known (Langley, 1977; Laurie-Ahlberg & Weir, 1979; Hill, 1981). Chakravarti *et al.* (1984) presented a method for estimating C that is appropriate for nucleotide data, but the statistical properties of their estimate are not known (Weir & Hill, 1986). Hudson & Kaplan (1985) proposed a method of estimating C with nucleotide data, but the method is difficult to apply, requiring simulations even to obtain an estimate, and the error bounds on the estimate are wide. In summary, no estimator of C is known that is well characterized statistically. In this note I present a new estimator of C which is relatively easy to calculate and that is appropriate for DNA sequence and restriction map data. Analytical results concerning the statistical properties of this estimate are not obtained, but Monte Carlo simulation results are presented which characterize the statistical properties of the estimator.

The estimator is based on the statistic S_k^2, the variance of the number of site differences between pairs of sequences in a sample. This quantity was first suggested by Sved (1968) as a measure of multilocus association. Its use for that purpose has been examined by Brown, Nevo & Feldman (1980), and by Chakraborty (1981, 1984).

2. The estimator

Consider a sample of n gametes, labelled from 1 to n, each of which has been sequenced at a homologous region (a locus) that is m nucleotide sites long. Let k_{ij} denote the number of sites at which gamete i and gamete j differ at the locus. Let S_k^2 denote the variance of the sample distribution of k_{ij}:

$$S_k^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} (k_{ij} - \bar{k})^2 / n^2 \qquad (1)$$

where $\bar{k}$ $(= \Sigma\Sigma k_{ij}/n^2)$ is the average of all the k_{ij}, including $i = j$. Brown, Feldman & Nevo (1980) found that S_k^2 can also be written as a function of the pairwise linkage disequilibria between the sites:

$$S_k^2 = \sum_{j}^{m} h_j - \sum_{j}^{m} h_j^2 + 2 \sum_{j}^{m} \sum_{l>j}^{m} \sum_{i} \sum_{k} [2 p_{ji} p_{lk} D_{ik}^{jl} + (D_{ik}^{jl})^2], \qquad (2)$$

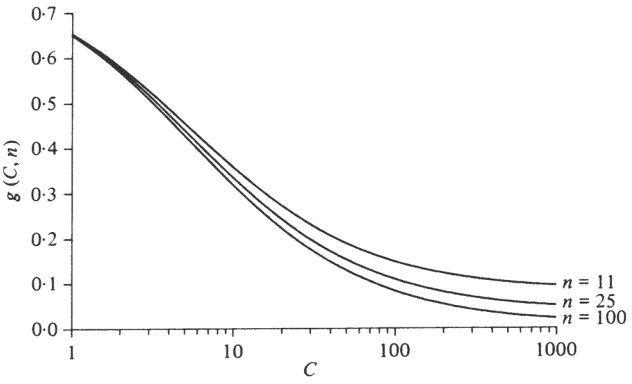

Fig. 1. The function $g(C,n)$ plotted as a function of C, for several values of n.

where p_{ji} is the sample frequency of the ith allele at site j, h_j $(=1-\Sigma p_{ji}^2)$ is the sample estimate of the heterozygosity at site j, and D_{ik}^{jl} is the sample estimate of the linkage disequilibrium between the ith allele at site j and kth allele at site l, that is,

$$D_{ik}^{jl} = g_{ik}^{jl} - p_{ji}p_{lk}$$

where g_{ik}^{jl} is the sample frequency of the gamete with allele i at site j and allele k at site l.

The first two sums on the right-hand side of (2) are just sums of single-locus quantities, and their expectations do not depend on the recombination rates between the sites. In the appendix the expectation of the quadruple sum on the right-hand side of (2) is calculated for an infinite-site neutral model described by Hudson (1983). It is assumed that the population is panmictic and at statistical equilibrium under the neutral model. In this model each of the m sites evolves according to an infinite-allele model with neutral mutation rate u/m. Also, if the m sites are labelled in order from 1 to m, the recombination rate between sites i and j is assumed to be $c|i-j|/(m-1)$. If m is large and u/m small the expectation of the quadruple sum in (2) is a function of $\theta(=4Nu)$, $C(=4Nc)$ and n. It is shown in the appendix that the expectation can be written as the product of θ^2 and a function of C and n, that is,

$$E(S_k^2 - \Sigma h_j + \Sigma h_j^2) = \theta^2 g(C,n). \tag{3}$$

This suggests the estimator, $\hat{C}$, which is defined as the solution of the following equation:

$$(S_k^2 - \Sigma h_j + \Sigma h_j^2)/\hat{\theta}^2 = g(\hat{C},n), \tag{4}$$

where $\hat{\theta} = \Sigma h_j[n/(n-1)]$, which is a nearly unbiased estimate of θ if $\theta/m \ll 1$. The function $g(C,n)$ is given in the appendix and is plotted in Fig. 1 for several values of n. The solution of equation (4) for any observed value of the left-hand side can be obtained approximately directly from Fig. 1, or the solution can be obtained easily numerically.

To investigate the statistical properties of $\hat{C}$, the Monte Carlo method of Hudson (1983) was used to generate random samples of gametes which were used to calculate $\hat{C}$. In Fig. 2 are shown estimates of the

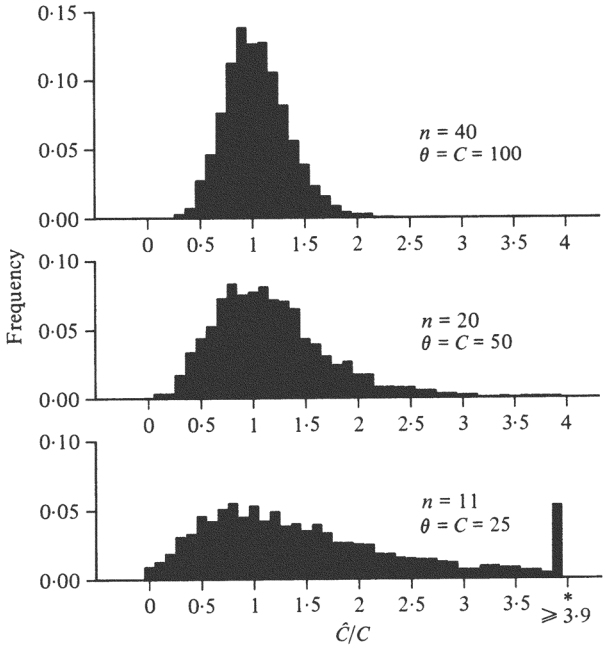

Fig. 2. The estimated distribution of $\hat{C}/C$ for three different sample sizes and values of θ and C. In each case, 2000 independent samples were generated to obtain the distribution.

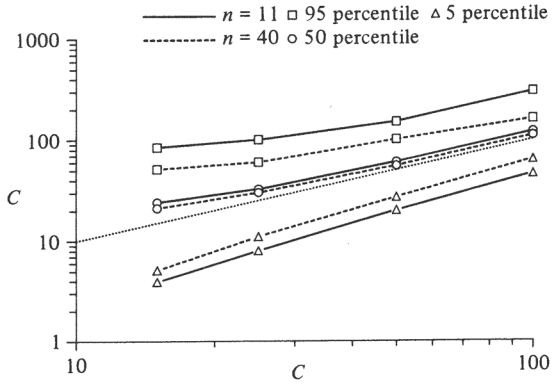

Fig. 3. Estimated percentiles of $\hat{C}$ as a function of C for samples of size 11 and 40. Estimates of the 5, 50 and 95 percentiles of the distribution of $\hat{C}$ were obtained at C equal to 15, 25, 50 and 100. The estimated median values are shown with circles. All estimates are based on 2000 samples.

distribution of $\hat{C}$ for three different cases. It is clear that, for $n = 11$ and $C = \theta = 25$, the estimate is likely to be very poor. Note especially the long and substantial tail to the right. With $n = 20$ and $C = \theta = 50$, the estimator is better but not likely to be very precise. In this case there is still a substantial probability that $\hat{C}$ differs from the true value by more than a factor of two, but the tail to the right is considerably reduced from the previous case. If the sample size is increased to 40 and the size of the region examined is increased so that $C = \theta = 100$, the distribution of $\hat{C}$ is fairly tightly centred about the true value, indicating that a fairly reliable estimate can be obtained in this case. In all three cases the mode of the distribution is near the true value.

https://doi.org/10.1017/S0016672308009610 Published online by Cambridge University Press

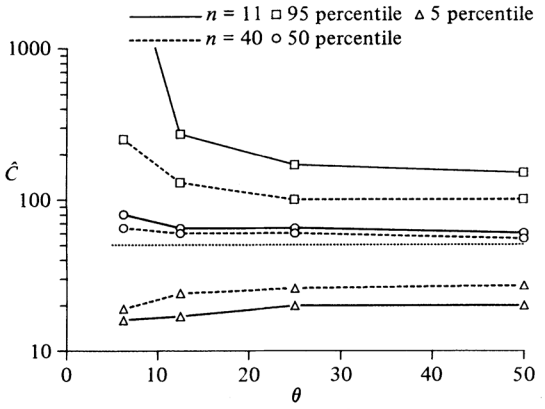

Fig. 4. Estimated percentiles of $\hat{C}$ as a function of θ for samples of size 11 and 40.

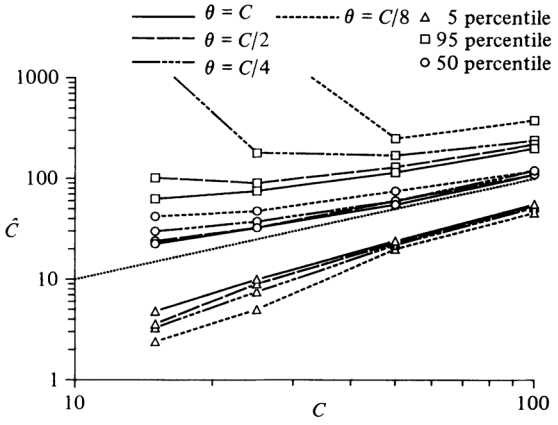

Fig. 5. Estimated percentiles of $\hat{C}$ as a function of C for samples of size 20 and four different levels of mutation, $\theta = C$, $C/2$, $C/4$ and $C/8$.

In Fig. 3, estimated percentiles of $\hat{C}$ are shown as a function of C, for the case of $C = \theta$ and for $n = 11$ and $n = 40$. The figure clearly shows that increasing the sample size from 11 to 40 substantially improves the estimate, especially for large values of C. Also, the estimate improves as C and θ increase, as one would expect. Note also that for these parameter values the median of the distribution of $\hat{C}$ is near the true value.

In Fig. 4 estimated percentiles of $\hat{C}$ are plotted as a function of θ, for C fixed at 50. For $n = 11$, the estimate deteriorates rapidly as θ decreases below 20. For $n = 40$, the estimate is better down to somewhat lower values of θ. For either sample size, the reliability of the estimator is essentially no better with $\theta = 50$ than with $\theta = 25$.

In Fig. 5, estimated percentiles of $\hat{C}$ are plotted as in Fig. 3, but for $n = 20$ and for several different mutation rates, namely $\theta = C$, $\theta = C/2$, $\theta = C/4$ and $\theta = C/8$. Clearly, lower mutation rates result in much poorer estimates if C is small enough, but if C is sufficiently large, then even if $\theta = C/8$ reliable estimates can be obtained.

3. An application

The *Drosophila melanogaster* sequence data obtained by Kreitman (1983) can be used to obtain an estimate of C, although it should be noted that his sample of flies was not a random sample. These data consist of a eleven sequences, 2·7 kilobases long, encompassing the *Adh* locus. There are 43 polymorphic nucleotide sites and 6 sites of length polymorphism. At two of the sites of length polymorphism, there are more than two length variants present in the sample. These two sites were ignored in the following calculations. For these data $n = 11$, and we calculate that $\hat{\theta} \sim 16$, $S_k^2 = 83{\cdot}6$ and $\hat{C} \sim 25$. Additional simulations show, for $n = 11$, $\theta = 16$ and $C = 80$, that the probability of $\hat{C}$ being less than or equal to 25 is about 0·025. We conclude that C is very likely less than 80. Since the tail of the distribution of $\hat{C}$ is so large for small values of C, no small value of C is incompatible with the observation $\hat{C} = 25$. Using the same data but a different method, Hudson & Kaplan (1985) estimated C to be approximately 35, and they were able to conclude that C was likely to be between 5 and 150. Thus their estimate is similar to ours, their upper bound on C is higher, but they are able to establish a lower bound as well.

Note that we can estimate c/u by $\hat{C}/\hat{\theta}$, and that for the Kreitman data this is approximately $25/16 \sim 1{\cdot}6$. This ratio can be estimated from completely independent empirical data. The average recombination rate per base pair in *D. melanogaster* females has been estimated to be $1{\cdot}7 \times 10^{-8}$ (Chovnick, Gelbart & McCarron, 1977). Since there is essentially no recombination in males and the sequences are 2·7 kilobases long, we estimate

$$c \simeq (1{\cdot}7 \times 10^{-8})(0{\cdot}5)(2{\cdot}7 \times 10^{3}) = 2{\cdot}3 \times 10^{-5}$$

for the *Adh* region sequenced by Kreitman. (Incidentally, we can with this information estimate N by $\hat{C}/4c \sim 3 \times 10^{5}$.) The neutral mutation rate has been estimated for a variety of organisms to be in the range 2×10^{-9} to 5×10^{-9} mutations per base pair per *year* (Li, Lou & Wu, 1985). If *D. melanogaster* averages four generations a year (probably an underestimate), then taking the neutral mutation rate to be 4×10^{-9}, $u \simeq (2{\cdot}7 \times 10^{3})\,4 \times 10^{-9}/4 = 2{\cdot}7 \times 10^{-6}$, and $c/u = (2{\cdot}3 \times 10^{-5})/(2{\cdot}7 \times 10^{-6}) \simeq 8$. In this calculation we have assumed that all sites mutate at the rate of 4×10^{-9}, including the 765 sites that code for protein. This is likely to be an overestimate of the average mutation rate of these sites. Given the generation time and the mutation rate that we have assumed, c/u could very plausibly be more than twice our estimate of eight. This estimate contrasts sharply with our estimate $\hat{C}/\hat{\theta} \simeq 1{\cdot}6$. Both of these estimates of c/u are subject to considerable error, but these calculations certainly suggest a problem. The problem may be due to the fact that our estimates of $\hat{C}$ and $\hat{\theta}$ rely on the assumption of a panmictic population at statistical

https://doi.org/10.1017/S0016672308009610 Published online by Cambridge University Press

equilibrium under the neutral model. There is now strong evidence that the molecular variation at the *Adh* locus of *D. melanogaster* is not compatible with this assumption, and that the level of polymorphism in the *Adh* locus is greater than would be expected under the neutral model (Kreitman & Aguadé, 1986*b*; Hudson, Kreitman & Aguadé, 1987). It is interesting to speculate that other regions of the *D. melanogaster* genome may not show such deviations from neutrality and may exhibit much larger c/u ratios, more in line with our expectations. It would also be very interesting to estimate the c/u ratio in species with long generation times such as humans. In these species the neutral mutation rate per *generation* may be considerably greater than in Drosophila but the recombination rate per generation may not be much different, in which case we would expect to see much smaller c/u ratios in these long-generation species than in short-generation species such as *D. melanogaster*.

4. Discussion

The estimator $\hat{C}$ is relatively easy to calculate and is a reliable estimator if a large enough data set is obtained. Unfortunately, such large data sets may require prohibitively large research efforts. We have seen that even with eleven sequences each 2·7 kilobases long, the estimate is not likely to be very precise. It appears that a sample four times as large and sequences four times as long would be needed to obtain a reliable estimate of C in *Drosophila melanogaster*. With current sequencing methods it appears that such large data sets are unlikely to be obtained. Intensive restriction-site mapping may be a more efficient method for obtaining information about C. However, since restriction-site mapping typically detects only a small fraction of the variability that is present at the sequence level, the ratio C/θ is made effectively larger. As shown in Figs. 4 and 5, if C/θ is more than four, poor estimates of C are likely, except when C is quite large. As discussed in the previous section, C/θ may be larger than four even with sequence data. If restriction mapping techniques made it possible to examine much longer regions of the genome, the larger C/θ ratio might not prevent good estimates from being obtained. Kreitman & Aguadé (1986*b*) have recently examined the same 2·7-kilobase segment that was discussed above in 87 lines using a battery of four-cutter restriction enzymes. With this technique they were able to detect approximately 20% of the polymorphisms that would have been detected by complete sequencing, resulting in an effective θ of about three. Even with the use of many four-cutter enzymes as in this case, if C equals 25 the ratio of C to the effective θ is about 8, and our results in Fig. 5 (for $n = 20$) suggest that one would have to examine a region more than four times as long to obtain reasonably good estimate.

In species with longer generation times, such as humans, the ratio of C/θ may be smaller and the use of restriction maps may be more efficient. Since such species may typically have smaller effective population sizes, it may be necessary to examine very long segments of DNA to obtain good estimates of C. For example, consider the question of how long a region must be examined in humans to make C equal to 100, a value which Figs. 3, 4 and 5 indicate is necessary to obtain a good estimate. If we take the effective population size of humans to be 10^4 and the recombination rate to be 2×10^{-8} per base pair, ($C = 100$) corresponds to about 100 kilobases. With this large a region, to obtain a reasonably reliable estimator of C one must still use a large number of restriction enzymes, so that the effective θ is at least $C/8 = 12{\cdot}5$ (see Fig. 5). If the effective θ is 12·5 the number of polymorphic restriction sites in a sample of 20 has expectation equal to approximately 12·5* log (20) $\sim$ 37. To summarize, these calculations suggest that, in order to obtain a reliable estimate of C in the human population, one must examine about 100 kilobases of DNA, in say twenty individuals, with enough different restriction enzymes so that about 40 restriction-site polymorphisms are observed.

Despite these difficulties, as demonstrated in Section 3, useful approximate estimates can be obtained with available data. More work needs to be done to evaluate the sensitivity of the estimates to departures from the equilibrium neutral model. For example, it would be useful to know how recent bottlenecks, population expansions or population subdivision would affect the estimates. It is also important that methods of estimating C such as that used by Chakravarti *et al.* (1984) be thoroughly investigated. Though some improvements of the method of Chakravarti *et al.* (1984) could certainly be made (Weir & Hill, 1986), the method uses information on the distances between the polymorphic sites, and therefore it may provide better estimates with smaller data sets than a method based on S_k^2.

References

Brown, A. H. D., Feldman, M. W. & Nevo, E. (1980). Multilocus structure of natural populations of *Hordeum spontaneum*. *Genetics* **96**, 523–536.

Chakraborty, R. (1984). Detection of nonrandom association of alleles from the distribution of the number of heterozygous loci in a sample. *Genetics* **108**, 719–731.

Chakraborty, R. (1981). The distribution of the number of heterozygous loci in an individual in natural populations. *Genetics* **98**, 461–466.

Chakravarti, A., Buetow, K. H., Antonarakis, S. E., Waber, P. G., Boehm, C. D. & Kazazian, H. H. (1984). Nonuniform recombination within the human β-globin gene cluster. *American Journal of Human Genetics* **36**, 1239–1258.

Chovnick, A., Gelbart, W. & McCarron, M. (1977). Organization of the Rosy locus in *Drosophila melanogaster*. *Cell* **11**, 1–10.

Ewens, W. J. (1979). *Mathematical Population Genetics*. New York: Springer-Verlag.

Hill, W. G. (1981). Estimation of effective population size from data on linkage disequilibrium. *Genetical Research* **38**, 209–216.
Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology* **23**, 183–201.
Hudson, R. R. (1985). The sampling distribution of linkage disequilibrium under an infinite allele model without selection. *Genetics* **109**, 611–631.
Hudson, R. R. & Kaplan, N. L. (1985). Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**, 147–164.
Hudson, R. R., Kreitman, M. & Aguadé, M. (1987). A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**, 153–159
Kreitman, M. (1983). Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. *Nature* **304**, 412–417.
Kreitman, M. & Aguadé, M. (1986*a*). Genetic uniformity in two populations of *Drosophila melanogaster* as revealed by filter hydridization of four-nucleotide-recognizing restriction enzyme digests. *Proceedings of the National Academcy of Sciences, USA* **83**, 3562–3566.
Kreitman, M. & Aguadé, M. (1986*b*). Excess polymorphism at the *Adh* locus in *Drosophila melanogaster*. *Genetics* **114**, 93–110.
Langley, C. H. (1977). Nonrandom associations between allozymes in natural populations of *Drosophila melanogaster*. In Lecture Notes in Biomathematics, 19, *Measuring Selection in Natural Populations* (ed. F. B. Christiansen and T. M. Fenchell), pp. 265–273. New York: Springer-Verlag.
Laurie-Ahlberg, C. & Weir, B. S. (1979). Allozyme variation and linkage disequilibrium in some laboratory populations of *Drosophila melanogaster*. *Genetics* **92**, 1295–1314.
Li, W.-H., Luo, C.-C. & Wu. C.-I. (1985). Evolution of DNA sequences, in *Molecular Evolutionary Genetics* (ed. R. J. MacIntyre). New York: Plenum
Ohta, T. (1980). Linkage disequilibrium between amino acid sites in immunoglobulin genes and other multigene families. *Genetical Research* **36**, 181–197.
Strobeck, C. & Morgan, K. (1978). The effect of intragenic recombination on the number of alleles in a finite population. *Genetics* **88**, 829–844.
Sved, J. A. (1968). The stability of linked systems of loci with small population size. *Genetics* **59**, 543–563.
Tajima, F. (1983). Evolutionary relationships of DNA sequences in finite populations. *Genetics* **105**, 437–460.
Weir, B. S. & Hill, W. G. (1980). Effect of mating structure on variation in linkage disequilibrium. *Genetics* **95**, 477–488.
Weir, B. S. & Hill, W. G. (1986). Nonuniform recombination within the human β-globin gene cluster. *American Journal of Human Genetics* **38**, 776–778.

Appendix

We outline here a method for obtaining the expectation of the quantity $H = S_k^2 - \Sigma h_j + \Sigma h_j^2$. We start by considering an m-site model where each site evolves according to an infinite-alleles model with mutation rate per generation of u/m. We assume that the recombination rate between site j and site l is $c_{jl} = c/l-j|/(m-1)$. For u and c small, the mutation rate for all m-sites together is u, and the recombination rate between the two most distant of the sites is c. We let C denote $4Nc$, as before, and we let C_{jl} denote $4Nc_{jl}$. We also define the sample identity coefficients:

$$\hat{\Phi}_{jl} = E(\Sigma\Sigma(g_{ik}^{jl})^2) \quad \text{and}$$
$$\hat{\Delta}_{jl} = E(\Sigma\Sigma(p_{ji}^2 p_{lk}^2)),$$

where g_{ik}^{jl}, p_{ji} and p_{lk} are as defined earlier in the body of this note.

As Brown, Feldman & Nevo (1980) pointed out, it is clear from (2) that H can be written as:

$$H = 2\sum_j \sum_{l>j} \sum_i \sum_k [(g_{ik}^{jl})^2 - p_{ji}^2 p_{lk}^2].$$

So obviously,

$$E(H) = 2\sum_j \sum_{l>j} [\hat{\Phi}_{jl} - \hat{\Delta}_{jl}]. \tag{A 1}$$

The quantity $\Phi_{jl} - \hat{\Delta}_{jl}$ is essentially what Ohta (1980) referred to as the identity excess. Therefore our estimator is actually based on the identity excess divided by an estimate of θ^2. The sample identity coefficients on the right-hand side of (A 1) are known under the neutral model, and as shown by Hudson (1985), can be written as:

$$\hat{\Phi}_{jl} = (1-1/n)\Phi_{jl} + 1/n, \tag{A 2}$$

and

$$\hat{\Delta}_{jl} = n_3 \Delta_{jl}/n^3 + 2n_2(\Phi_j + 2\Gamma_{jl})/n^3 + 2n_1(2\Phi_j + \Phi_{jl})/n^3 + 1/n^2, \tag{A 3}$$

where $n_i = (n-i)(n-i+1)\ldots(n-1)$ and $\Phi_j = 1/(1+\theta/m)$, and where Δ_{jl}, Γ_{il} and Φ_{jl} are population identity coefficients whose values are known. Using the formulas for these population identity coefficients given by Strobeck & Morgan (1978) and assuming that the recombination rate between sites j and l is $c_{jl} = c|j-l|/(m-1)$ we find after a good deal of algebra that:

$$E(H) = 2\sum_j \sum_{l>j} [(\theta^2/m^2) f(C_{jl}) + O(1/m^3)]$$
$$= 2\sum_{i=1}^{m} [(\theta^2/m^2) f(iC/m)(m-i) + O(1/m^2)], \tag{A 4}$$

where

$$f(z) = [(z+14) + z(z+12)/n - (z+2)(z+13)/n^2 + 2(z+6)/n^3]/D,$$

with $D = z^2 + 13z + 18$. This result can also be obtained from the results of Weir & Hill (1980). The limit as m tends to infinity of the right-hand site of (A 4) is just a Riemann integral, that is, for large m:

$$E(H) \simeq 2(\theta^2/C^2)\int_0^C f(z)(C-z)\,dz$$
$$= \theta^2 g(C, n)$$

https://doi.org/10.1017/S0016672308009610 Published online by Cambridge University Press

where

$$g(C,n) = (2/C^2)\int_0^c f(z)\,(C-z)\,dz$$

$$= (2/C^2)\{(-C+(C\text{-}1)\,I_1+2(7C+9)\,I_2)$$
$$+(C^2/2+C+(5-C)\,I_1-18(C+1)\,I_2)/n$$
$$+(-C^2/2+2C-2(C+9)\,I_1-4(2C+9)\,I_2)/n^2$$
$$+(-2C+2(C+7)\,I_1+12(C+3)\,I_2)/n^3\},$$

$$I_1 = \int_0^c \frac{z}{z^2+13z+18}\,dz$$

$$= \tfrac{1}{2}\log[(C^2+13C+18)/18]-\tfrac{13}{2}I_2,$$

and

$$I_2 = \int_0^c \frac{dz}{z^2+13z+18}$$

$$= \frac{1}{\sqrt{97}}\log\frac{(2C+13-\sqrt{97})(13+\sqrt{97})}{(2C+13+\sqrt{97})(13-\sqrt{97})}$$

https://doi.org/10.1017/S0016672308009610 Published online by Cambridge University Press