ON THE FREQUENCY OF UNDETECTABLE RECOMBINATION EVENTS

J. CLAIBORNE STEPHENS

Center for Demographic and Population Genetics, University of Texas at Houston, P.O. Box 20334, Houston, Texas 77225

Manuscript received October 3, 1985
Accepted December 5, 1985

ABSTRACT

Simple analytical results show that many recombination events occur in such a way as to have no effect on the resultant DNA sequence. The proportion of these undetectable events depends on the population size, mutation rate and recombination rate and is quite large for reasonable values of these quantities. Efforts to estimate recombination rates and frequencies directly from DNA sequence data must, therefore, take this undetectable fraction into account.

RECENT advances in molecular technology have allowed population geneticists to evaluate genetic variation at the most fundamental level—that of the DNA sequence. Population surveys using restriction site or nucleotide sequence data have provided insights into a variety of fundamental issues, spanning interests ranging from biochemistry and molecular biology to evolutionary theory. One particular concern is the importance of recombinational events at the resolution of several kilobases of DNA sequence: what is the rate and frequency of recombinational events, and what impact does recombination have on our ability to reconstruct the evolutionary history of a sample of DNA sequences?

HUDSON and KAPLAN (1985) have considered the latter question from the standpoint of detectability of recombination events. They studied the theoretical sampling distribution of the number of recombination events that have occurred during the history of a sample of DNA sequences. Furthermore, they used computer simulations to compare the known number of recombination events with the number inferred by a detection technique based on parsimony, with the result that only a small fraction of known recombination events were inferred. If the only available information consists of a sample of DNA sequences, the number of recombination events that have occurred in this sample can be divided into two categories, say I and II. Category I includes recombination events that do not result in any observable effect on the DNA segments; such effects are undetectable by any detection algorithm. Recombination events in category II affect the DNA sequence and, hence, are potentially

detectable. In this note I should like to show that the frequency of category I recombinations is very high unless sequence diversity is extremely high.

The model of recombination considered here is that of single crossovers, excluding double crossovers or gene conversions. Single crossovers are physical exchanges of genetic information between homologous nonsister chromatids and will be counted as recombination events, whether or not they result in any observable change in DNA sequence. This is, in fact, the convention used by HUDSON and KAPLAN (1985), as seen by noting that their equation (5) and table 1 are independent of the mutation rate.

Note, however, that for a recombination event to be potentially detectable (in category II), the crossover point must occur between two flanking segregating sites. Obviously, a crossover between identical sequences or between sequences that differ at a single site is undetectable. Even if a pair of sequences differ by two or more sites, a crossover in the region flanking the segregating sites would be undetectable. Hence, with information about the number of site differences between a pair of sequences and the location of such differences within the sequence, we can estimate the proportion of events that would be in category I.

The number of nucleotide differences (d) between a random pair of DNA sequences is related to the quantity $\Theta = 4Nv = 4Nn\mu$, where N is the effective population size, v is the mutation rate per sequence, n is the number of nucleotides in the sequence and μ is the mutation rate per nucleotide site. In a population without recombination under the usual assumptions of constant population size and no selection, the probability distribution of d is approximately geometric (WATTERSON 1975):

$$P(d = k) = \Theta^k/(\Theta + 1)^{k+1}. \quad (1)$$

If the recombination rate is very small, we may assume that (1) is still applicable. In random mating diploid organisms, (1) represents the relative frequency of genotypes where k nucleotides differ between the two homologous DNA segments.

As mentioned above, intragenic recombination corresponding to genotypic classes with $d = 0$ and $d = 1$ has no detectable effect, whereas for genotypic classes with $d \geqq 2$, recombination will be detectable only if the breakpoint occurs within the region spanned by the sites where nucleotide differences exist. The relative proportions of the latter events will therefore depend on the positions of these heterozygous nucleotide sites.

In a model of n independent, equivalent nucleotide sites, the expected distance spanned by d randomly chosen sites is $(d - 1)(n + 1)/(d + 1)$, (STEPHENS 1985). Therefore, in a heterozygote whose alleles differ by $d \geqq 2$ sites, the average proportion of the nucleotide sequence in which a recombination would not be detectable is $1 - (d - 1)(n + 1)/n(d + 1) = 2/(d + 1)$, assuming that n is reasonably large. Hence, the proportion of recombination events which are undetectable is obtained as the sum of the probabilities in (1), weighted by the appropriate factor [*i.e.*, 1, 1 and $2/(d + 1)$, when $d \geqq 2$]. That is, the expected

TABLE 1

Proportion of undetectable recombination events

	Θ							
d	1	5	10	15	20	50	100	200
0	0.50	0.17	0.09	0.06	0.05	0.02	0.010	0.005
1	0.25	0.14	0.08	0.06	0.05	0.02	0.010	0.005
2+	0.14	0.24	0.22	0.19	0.17	0.10	0.062	0.038
Total	0.89	0.55	0.39	0.31	0.26	0.14	0.082	0.048

proportion of events in category I is

$$E(\mathrm{I}) = 2[\log_e(1 + \Theta)]/\Theta - 1/(1 + \Theta). \quad (2)$$

Some representative values are shown in Table 1.

It is clear, although somewhat surprising, that even for relatively large values of Θ, a substantial fraction of recombination events would occur in such a way that they could not be detected. It is also clear that a substantial proportion of the undetectable events occur in genotypic classes with $d \geqq 2$, especially when Θ and sequence diversity are high.

As recombination rates increase from zero, the geometric distribution in (1) should be replaced by a more bell-shaped distribution (R. R. HUDSON, personal communication). The Poisson distribution

$$P(d = k) = \exp(-\Theta)\Theta^k/k! \quad (3)$$

is the limiting distribution, corresponding to free recombination between sites. Using (3) in place of (1), (2) becomes

$$E(\mathrm{I}) = 2[1 - \exp(-\Theta)]/\Theta - \exp(-\Theta). \quad (4)$$

The totals corresponding to those in Table 1 become 0.90, 0.39, 0.20, 0.13, 0.10, 0.04, 0.02 and 0.01. Thus, it appears that for $\Theta < 20$, at least 10% of all recombination events will be undetectable. If the sequences are taken from a random-mating natural population, $4N\mu$ is almost always smaller than 0.02 (NEI 1983), so that the only way to obtain $\Theta > 20$ is to increase n by sequencing several kilobases.

The present results indicate that of all recombination events occurring within a population, a substantial fraction will be undetectable. In this regard, the proportion of undetected events in HUDSON and KAPLAN'S (1985) simulations was 76% or greater, even for their largest Θ-value (30) and smallest recombination rate. Events in category I constitute the lower bound on the number that is actually undetected in any given analysis. Other uncounted recombination events may arise from category II, including (1) redundant recombination events, such as those creating alleles which are already present in the population, and (2) inefficiency inherent in the method of detection. Events arising from the former are not liable to be an important source of error unless recombination rates are relatively large and gene diversity (as measured by Θ) is relatively low.

It is clear that any attempt to estimate rates and/or frequencies of recombination directly from DNA sequence data must allow for the undetectable fraction of events. Efforts to enumerate recombination events and to reconstruct evolutionary trees in the presence of recombination will need to focus on category II events.

This work benefited greatly from discussions with my colleagues M. NEI, R. CHAKRABORTY and R. HUDSON. This research was supported by grants from the National Institutes of Health (GM 20293) and the National Science Foundation (BSR 83115) to M. NEI.

LITERATURE CITED

HUDSON, R. R. and N. L. KAPLAN, 1985 Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics **111:** 147–164.

NEI, M., 1983 Genetic polymorphism and the role of mutation in evolution. pp. 165–190. In: *Evolution of Genes and Proteins,* Edited by M. NEI and R. K. KOEHN. Sinauer, Massachusetts.

STEPHENS, J. C., 1985 Statistical methods of DNA sequence analysis: detection of intragenic recombination or gene conversion. Mol. Biol. Evol. **2:** 539–556.

WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Pop. Biol. **7:** 256–276.

Communicating editor: B. S. WEIR