Proc. Natl. Acad. Sci. USA
Vol. 78, No. 6, pp. 3748–3750, June 1981
Genetics

Estimation of genetic variation at the DNA level from restriction endonuclease data

(genetic polymorphism/population genetics/heterozygosity/recognition sequence)

W. J. Ewens*, Richard S. Spielman†, and Harry Harris†

Departments of *Biology and †Human Genetics, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104

Contributed by Harry Harris, March 2, 1981

ABSTRACT We consider the estimation of the genetic variation in a natural population when the data are obtained by the use of restriction endonucleases. Under the restriction endonuclease technique, a particular DNA segment is considered and cut wherever a recognition sequence appropriate to the endonuclease occurs. We consider data generated when a random sample of homologous DNA segments is treated in this way with one or a battery of restriction endonucleases. The numbers and sizes of the fragments that result indicate the locations and the frequencies of the recognition sequence (or, with a battery of restriction endonucleases, of each recognition sequence). These frequencies in the sample form the basis for an estimate of the amount of genetic variation in the population.

The data we consider arise when a survey is made of homologous segments of DNA in a random sample of individuals from some population. By using the restriction endonuclease technique, each segment is cut wherever a given nucleotide recognition sequence occurs in the sequence. The numbers and sizes of the resulting fragments clearly give some information on the genetic heterogeneity of the population. If all the DNA segments are cut into the same number of fragments and the sizes of these fragments are the same from one member of the sample to another, we have evidence of genetic homogeneity in the population whereas, if the numbers and sizes of the fragments differ among the members of the sample, we have evidence of genetic heterogeneity. Our aim is to quantify these notions in a more precise fashion.

Definitions

It is assumed that the sizes of the various fragments can be used without error to determine the locations of the recognition sequences in the DNA segments. When this determination is made, we define a "cleavage site" as a block of consecutive nucleotide sites (usually six or four) for which at least one DNA segment in the sample has the recognition sequence of the restriction endonuclease(s). We say a recognition sequence is "monomorphic in the sample" at such a block if all DNA segments in the sample have a recognition sequence in this block (and are thus all cleaved, or cut, at this block). Similarly, a recognition sequence is defined to be "monomorphic in the population" at any block if all the DNA in the population has the recognition sequence at that block. (Of course, from sample data only, we can never be certain when this occurs.)

To clarify the points at issue we make two simplifications to introduce the argument. The first is that only one restriction endonuclease is used and that it recognizes a nucleotide sequence six bases long. The second concerns those cleavage sites that are not monomorphic in the sample at a given block (so that some, but not all, DNA segments have the recognition sequence at this block). Because of the large degree of homology between DNA segments from different individuals, we assume that the differences between the DNA segments are due to a difference of one nucleotide at one site only among the six sites in the block. For purposes of this paper, we ignore variation due to changes involving more than one nucleotide. More general results for a battery of endonucleases are given toward the end of this paper.

The data used in restriction endonuclease work to estimate genetic variation concern only cleavage sites. The reason for this is that if a recognition sequence does not occur in any individual at some arbitrary block of six sites, we have no knowledge concerning which of the 4095 alternative sequences do occur. Thus, any such block is practically useless for the purpose of assessing the amount of genetic variation at that block. The fact that conclusions are based on a special subset of blocks (i.e., where the recognition sequence occurs at least once in the sample) introduces a bias into standard estimation procedures, and we argue below that application of standard or intuitive methods leads to estimates of genetic variation that are biased by a factor of two. To discuss this more fully, we must consider the population genetics theory applicable to this situation.

Population genetics theory

In the classical population genetics theory of the frequencies of two alleles (A_1 and A_2) at a given gene locus, the frequency (x) of A_1 is a random variable having, at stationarity [under a neutral Wright–Fisher model (1)], the probability distribution

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{b-1}(1-x)^{a-1}, \qquad [1]$$

where $a = 4Nv_1$, $b = 4Nv_2$, N is the population size, and v_1 (v_2) is the mutation rate from A_1 (A_2) to A_2 (A_1). From this, it follows that if a sample of n ($n \ll N$) genes is taken at random from the population in any generation and it is assumed that a and b are both small (0.1 or less), the probability (P_0) that there are no A_1 genes in the sample and the probability (P_n) that all n genes in the sample are A_1 are given by

$$P_0 \approx an^{-b}/(a+b), \qquad P_n \approx bn^{-a}/(a+b), \qquad [2]$$

respectively, provided n is moderately large (at least 30). These formulae can be taken over and applied at the nucleotide site level, provided appropriate values are given for a and b, and from now on the analysis proceeds at that level. Assume that any base mutates with probability u, with equal probability of mutating to each of the other three base types. Fix attention on a specified nucleotide type, and identify this with A_1 and the remaining three nucleotide types collectively with A_2. Then, from Eq. 2, the probabilities (P_0 and P_n) that, at a given site, this nucleotide type is seen 0 or n times in the sample of n DNA segments are

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

$$P_0 \approx \frac{3}{4} n^{-\theta/3}, \quad P_n \approx \frac{1}{4} n^{-\theta}, \qquad [3]$$

respectively, where $\theta = 4Nu$. Standard theory also shows that if two homologous DNA segments are compared, the probability that, at any site, they exhibit different nucleotides is

$$\text{prob(heterozygosity)} \approx \theta. \qquad [4]$$

Further, the probability that, at any site, all $2N$ bases in the entire population are of the same nucleotide type is

$$\text{prob(population site monomorphism)} \approx (2N)^{-\theta}. \qquad [5]$$

The probability of population heterogeneity at any site is then, immediately,

$$\text{prob(population site heterogeneity)} \approx 1 - (2N)^{-\theta} \qquad [6]$$

and, similarly, the probability of heterogeneity at any site in a sample of n DNA segments is

$$\text{prob(sample site heterogeneity)} \approx 1 - n^{-\theta}. \qquad [7]$$

We emphasize that Eqs. 2–7 hold when n is moderate and θ is small and when selective neutrality obtains. Eqs. 4–7 originally arose for the "infinite alleles" model but can be applied here to a sufficient approximation because of the assumed smaller value of θ.

Application to restriction endonucleases

Our main aim, from restriction endonuclease data, is to estimate the probability of sample monomorphism (that, at any site, all individual DNA segments in the sample have the same nucleotide type). Associated aims are to estimate the probability of population monomorphism, of population heterozygosity, and of other similar measures of the amount of genetic variation in the population. Suppose then that, in our sample of n DNA segments, there are m cleavage sites and that, of these cleavage sites, k are *not* monomorphic in the sample (so that, at the remaining $m - k$ cleavage sites, all DNA segments in the sample are cut at each site). We interpret these data as indicating that, of the $6m$ sites surveyed, $6m - k$ are monomorphic in the sample and k are not.

The standard and intuitive estimate of the probability that a site taken at random in the population is monomorphic in the sample is then $1 - (k/6m)$. This estimate, however, is biased. In terms of the theory and notation developed in the previous section, our aim is to estimate $4P_n$ whose value, from Eq. 3, is $n^{-\theta}$. However, because the recognition sequence must occur in at least one individual in the sample for a particular cleavage site to be identified, $1 - (k/6m)$ is *not* an estimator of $4P_n$. Instead, it estimates the conditional probability that a nominated nucleotide, which occurs at a given site at least once in the sample, occurs at that site in all DNA segments in the sample. In other words $1 - (k/6m)$ is an estimator not of $4P_n$, but of $P_n/(1 - P_0)$, which can be written, from Eq. 3, as

$$P_n/(1 - P_0) \simeq n^{-\theta}/(4 - 3n^{-\theta/3}). \qquad [8]$$

For small values of θ and moderate values of n, we have

$$4P_n \simeq 1 - \theta \ln n, \qquad [9]$$
$$P_n/(1 - P_0) \simeq 1 - 2\theta \ln n.$$

To this degree of accuracy, $1 - (k/6m)$ is an estimator of $1 - 2\,\theta \ln n$, whereas we require an estimator of $1 - \theta \ln n$. Therefore, we should estimate the probability of sample monomorphism (on a nucleotide-site basis) by

$$1 - (k/12m). \qquad [10]$$

Correspondingly, the estimate of the probability of genetic variation in the sample (on a site basis)—i.e., of the fraction of sites in the sample at which two or more nucleotide types appear—is not $k/6m$ but rather $k/12m$. The intuitive estimator $k/6m$ used in the literature (see, e.g., ref. 2) differs from the correct value by a factor of two.

It may be noted from Eqs. 3–5 that the single parameter θ determines all probabilities of interest. From Eq. 9, we estimate θ by

$$\hat{\theta} = k/(12m \ln n). \qquad [11]$$

Substituting the value of $\hat{\theta}$ for θ in Eq. 4, we can now estimate the probability of heterozygosity (on a site basis) and, if a reasonable extrinsic estimate of the population size N is available, the probability of population monomorphism (on a nucleotide-site basis) by using Eq. 5.

The above theory assumes a single restriction endonuclease recognizing a sequence six bases long. The theory remains essentially unchanged if a battery of restriction endonucleases is used, provided all members of the battery recognize a six-base sequence. (We now interpret m as the total number of cleavage sites—i.e., as the total number of sites where a cut occurs in at least one member of the sample when one or another restriction endonuclease is used.) Suppose now some members in the battery recognize four-base sequences and some recognize six. Let m_4 be the total number of cleavage sites for all restriction endonucleases recognizing four-base sequences and, of these, let k_4 be the number of cleavage sites where sample monomorphism does *not* obtain (i.e., for which not all DNA segments in the sample are cut). Define m_6 and k_6 correspondingly for six-base recognition sequences. Then, the extension of Eq. 11 is

$$\hat{\theta} = (k_4 + k_6)/[(8m_4 + 12m_6) \ln n]. \qquad [12]$$

Correspondingly, the estimate of the fraction of sites in the DNA segments considered at which two or more nucleotide types appear in the sample is

$$(k_4 + k_6)/(8m_4 + 12m_6) \qquad [13]$$

rather than the intuitive value $(k_4 + k_6)/(4m_4 + 6m_6)$.

Data analysis

To illustrate the above theory, we consider the data of Jeffreys (2), who used a battery of eight restriction endonucleases, seven recognizing sequences of length 6 and one recognizing a sequence of length 4. The data refer to segments of DNA containing various globin genes in humans; 60 individuals were studied. Technical problems prevented unambiguous decisions in all segments at two sites, but for purposes of illustrating the above theory, we ignore these difficulties and assume that the battery was applied successfully to all 120 homologous DNA segments so we put $n = 120$. There were seven cleavage sites for the restriction endonuclease having a four-site recognition sequence, and all 120 segments were cleaved at all seven sites. Thus, $m_4 = 7$ and $k_4 = 0$. There were 47 cleavage sites for restriction endonucleases having six-site recognition sequences and, of these, all 120 segments in the sample were cleaved at 44. Thus, $m_6 = 47$ and $k_6 = 3$. By using Eq. 13, we estimate the fraction of sites in the DNA segment having two or more nucleotide types represented to be

$$3/(56 + 564) \simeq 0.0048. \qquad [14]$$

By contrast, the biased "intuitive" estimate is $3/310 \simeq 0.0096$. Further, from Eq. **12**, our estimate of θ is

$$\hat{\theta} \simeq 0.0048/\ln 120 \simeq 0.001. \qquad [15]$$

From Eq. **4**, this is also our estimate of heterozygosity—i.e., of the probability that when two homologous DNA sequences are compared, they will show different nucleotide types at a given site. Assuming a population size of 10^7, we can also use the result for $\hat{\theta}$ to estimate the probability of population-site monomorphism—i.e., the probability that, at a given site, all gametes in the population have the same nucleotide type. By using Eq. **5** this probability is estimated to be

$$(2 \times 10^7)^{-0.001} \simeq 0.983. \qquad [16]$$

We can also use Eq. **7** to estimate the probability of heterogeneity at any site in a sample of 120 as

$$1 - (120)^{-0.001} \simeq 0.0048, \qquad [17]$$

and this confirms the value in Eq. **14.**

Discussion

As one of our aims is, in effect, to estimate the fraction of sites in the DNA segment considered at which, in the sample, only one nucleotide type occurs, it is perhaps natural that we use as data the fraction of cleavage sites for which all DNA segments in the sample are cleaved. From a statistical point of view, however, we can view our procedures as being an estimation of the fundamental parameter θ and, from this point of view, we should, for optimal estimation, consider all the data afforded by the experiment (and then use optimal—i.e., maximum likelihood—estimation procedures). The data we have ignored above are the numbers of cut and uncut DNA segments at those cleavage sites where not all segments are cut. However, the use of this information does not significantly alter the estimate (Eq. **15**) (unpublished results). Further, we show that difficulties arise with maximum likelihood techniques but that the procedure leading to Eq. **15** uses the bulk of the information concerning θ afforded by the experiment.

Other statistical problems concern standard errors and the dependence of our conclusions on the specific (Wright–Fisher) model adopted. It appears that if selective neutrality can be assumed, the estimate (Eq. **13**) is independent of the model used. Our estimation procedures for θ apply whether or not there is linkage disequilibrium between the various sites considered. To derive the variance of the estimate, however, it would be necessary to take into accout the possibility of linkage disequilibrium.

A further point of interest concerns the relationship between the "sites" θ (estimated in Eq. **15**) and the "genes" θ. In principle it should be possible, knowing the average number of nucleotides per gene, to extrapolate from the estimate of θ obtained here to the predicted value for an entire gene. However, a comparison between the two values is difficult for several reasons. It is likely that neither the well-established data for θ at the polypeptide level nor these first data for θ at the DNA level are representative of the entire genome and, furthermore, that the two kinds of data are sampling different "parts" of the genome. The Jeffreys data on cleavage site polymorphisms refer only to the DNA immediately adjoining the γ- and δ–β-globin genes and these polymorphisms exhibit certain features that exemplify the difficulties.

(*i*) All three polymorphic cleavage sites in Jeffreys' data are located in intervening sequences (introns), which might be expected *a priori* to be more variable than those parts of genes that are reflected in polypeptides.

(*ii*) The $^G\gamma$ and $^A\gamma$ sequences represent a gene duplication, and the polymorphisms detected by restriction endonucleases in each are apparently at homologous cleavage sites. Thus the polymorphisms detected are probably not statistically independent. They may, for example, each have been derived from the same original mutation.

(*iii*) As the entire DNA screened with a single probe represents very closely linked genetic material, it is likely that there is linkage disequilibrium among the cleavage sites.

(*iv*) As is well known, the polypeptide data, obtained primarily by electrophoresis, probably underestimate the frequency of amino acid substitutions by a factor of perhaps 3. In addition, nucleotide substitutions that do not result in an amino acid substitution are not detected by analysis of polypeptides.

(*v*) DNA not coding for structural gene products is not represented in the polypeptide data. To the extent that the polypeptide and DNA data refer to functionally different segments of genes, the two kinds of data may reflect quite distinct evolutionary processes.

This work was supported by National Institutes of Health Grants GM 21135, GM 20138, and GM 27018.

1. Ewens, W. J. (1979) *Mathematical Population Genetics* (Springer, Berlin), pp. 16–21.
2. Jeffreys, A. J. (1979) *Cell* **18**, 1–10.