Proc. Natl. Acad. Sci. USA
Vol. 75, No. 10, pp. 5094–5098, October 1978
Genetics

Estimation of expected number of rare alleles of a locus and calculation of mutation rate

(population genetics/Amerindian tribes/population structure)

EDWARD D. ROTHMAN* AND JULIAN ADAMS†

* Department of Statistics and Human Genetics and † Division of Biological Sciences and Department of Human Genetics, University of Michigan, Ann Arbor, Michigan 48109

Communicated by James V. Neel, July 10, 1978

ABSTRACT **An approach is described for the estimation of the number of rare variants in the population from the number in a sample drawn at random from the population. This quantity is used to derive an estimate of the mutation rate. The data required are the number of rare variants in the sample and the distribution of offspring within a population of well-defined size with little or no immigration. Application of this approach to data on 28 loci assayed in the Yanamamo, a tribe of South American Indians, yields an average mutation rate of $0.1 \sim 0.2 \times 10^{-5}$. Determination of this figure is subject to several assumptions concerning the nature of the rare variants and the structure of the population. Violation of these assumptions will generally result in the underestimate of the true mutation rate.**

A variety of approaches, both direct and indirect, have been used in the past to estimate mutation rate in human populations. Although direct approaches are most reliable, they are practical only for dominant or semidominant mutations and where the acquisition of extremely large samples is both possible and feasible. Consequently, indirect approaches have been used predominantly in the past because the data for such approaches are easier to acquire and often mutation rate estimates can be obtained whatever the phenotype of the heterozygote. The most common and most widely used indirect approach postulates an equilibrium between mutation rate and the effect of natural selection, and can be used for both dominant and recessive autosomal mutants as well as for sex-linked mutants. However, its utility is limited by a number of restrictive assumptions that are seldom met in practice (1).

A criticism that can be applied equally to the direct and indirect methods described above is that they can measure only one type of mutational event, mutation of the wild-type allele to a mutant form with a clearly defined phenotype (such as complete dysfunction). This clearly defines a subset of all the mutational events occurring at a given locus and, thus, any estimate obtained by these approaches must be considered to be an underestimate. Such estimates may be useful for the consideration of the impact of particular genetic traits, but they are not so valuable for a general consideration of the effect of mutations in human populations.

A completely different approach was suggested several years ago by Kimura and Ohta (2), and first applied by Neel (3). Modifications of this procedure have recently been proposed by Nei (4). This estimate is based on the following formula, which describes the number of different neutral alleles in a population in terms of their rate of production due to mutation and their rate of loss due to random events in a finite population. The relationship is

$$2N\mu = K/t_o, \qquad [1]$$

where N is the population size, μ is the mutation rate, K is the number of alleles in the population, and t_o is the expected number of generations an allele remains in the population. The utility of this estimate is limited by the ability to detect and distinguish genetic variants by such biochemical screening techniques as electrophoretic mobility. Although this approach will still only identify a subset of all mutational changes, the number of mutational events identified will clearly be larger than in the approaches described above. In addition, this approach can be easily modified to incorporate estimates of the expected proportion of mutations missed by the particular screening technique employed.

In spite of the appealing simplicity of this formula, there are two problems involved in its use for estimation of mutation rates in human populations. (*i*) The number of alleles in the population is generally unknown though the number of alleles in the sample may be easy to obtain. (*ii*) The expected time before extinction of an allele is difficult to estimate if the population has any structure. In this paper we estimate the expected number of alleles in a population from the observed number of alleles in a sample and, using this estimate, we describe a procedure for calculating mutation rates.

MODEL AND ESTIMATION PROCEDURE

A population of finite size will possess a characteristic number of alleles at a given locus, with a characteristic distribution of frequencies or numbers of copies of these alleles. The number of alleles present as singletons in any one generation can be described by the following simple equation,

$$Kg(1) = 2N\mu + K \sum_j g(j)P_{j1}, \qquad [2]$$

where $g(j)$ is the expected relative frequency of alleles present as j copies in the population and K is the expected number of alleles in the population. P_{ji} is the probability that an allele is present in i copies in the present generation given that it is present in j copies in the previous generation; that is,

$$P_{ji} = Pr[X_m = i | X_{m-1} = j], \qquad [3]$$

where X_m is the number of copies of an allele in generation m. We assume here only that an infinite series of alleles can be generated at this locus; namely, every new mutation is assumed to result in a novel allele. This is commonly known as an infinite alleles model. To use Eq. 2 to estimate mutation rate, it is necessary to assume also that the distribution of the relative frequencies of the alleles is constant from one generation to the next. It should be noted that this assumption is less restrictive than the assumption of a comprehensive equilibrium at which all population parameters are constant. For example, the relative allele frequencies may remain constant even though population size is increasing and the numbers of copies of the alleles are changing.

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.

Estimation of the mutation rate by Eq. 2 requires knowledge of $g(j)$, P_{ji}, and K, and we now describe procedures to obtain estimates of these parameters.

Estimation of P_{ji}. Although only P_{j1} is specified explicitly in Eq. 2, it will be necessary to obtain P_{ji} for calculation of $g(j)$ described in the next section. In general, estimation of the mutation rate will involve the determination of P_{ji} for $i,j = 1, 2, \ldots, 2N$, where N is the population size. It is clear that for a population of any appreciable size, the number of values of P_{ji} to be determined will become impossibly large. However, we take advantage here of the consideration that the majority of alleles will be present in very low frequency if $4N_e\mu \ll 1$, where N_e is the effective population number (5, 6). Thus, providing this inequality is true, we can focus attention only on the rare variants (2) in the population. This simplifies our estimation procedure considerably, as we only need to calculate the P_{ji} for i,j small. For the purposes of this treatment, we consider rare alleles to include the rare variants and private polymorphisms defined in the following paper (7), i.e., alleles whose frequency in the population is small compared to the population size (7).

Under some circumstances, for example, when detailed pedigree data are available for the population, it may be possible to determine P_{ji} directly. In the absence of such information, we must proceed indirectly and calculate P_{ji} from the offspring distribution.

If the distribution of offspring $A(x)$ for members of the population is known, then, assuming Mendelian inheritance, we can deduce the probability mass function for the number of copies produced by a given allele, $B(x)$, to be

$$B(x) = \sum_{r=x}^{\infty} \binom{r}{x} \left(\frac{1}{2}\right)^r A(r). \quad [4]$$

The probability generating function corresponding to $B(x)$ is

$$f(s) = \sum_{r=0}^{\infty} B(x)s^x \quad [5]$$

If we assume that the rare alleles are neutral in effect, the number of copies produced by each rare allele will be independent of each other, each with the same probability generating function $f(s)$. Then it follows by standard statistical procedures that

$$P_{ji} = Pr[X_m = i \,|\, X_{m-1} = j] = \text{coefficient of } s^i \in [f(s)]^j. \quad [6]$$

In most situations the probability generating function $f(s)$ will be unknown, since $A(x)$ will be unknown. However, it is possible to obtain reliable estimators for P_{ji} when a good estimate of $A(x)$ is available.

Estimation of $g(j)$. Although it is possible to determine the relative frequencies of the rare alleles in the sample from the data, we cannot obtain the relative frequencies $g(j)$ of the alleles in the population so simply. We are restricted to the following procedure, which uses the transition probabilities calculated earlier.

Eq. 2 describes the expected relative frequency of alleles present as singletons in the population. In a similar fashion, it is easy to see that the expected relative frequency of alleles with i copies in the population will be

$$\sum_{j=1}^{\infty} g(j)\, P_{ji} = g(i). \quad [7]$$

We can combine Eqs. 2 and 7 and write them in matrix form in the following manner.

$$\mathbf{G} = (I - P')^{-1}\,\mathbf{M} \quad [8]$$

where $\mathbf{M} = (2N\mu/K, 0, 0 \ldots . 0)$, $\mathbf{G}$ is the vector of relative rare allele frequencies ($g(j)$), and P is the matrix of transition probabilities.

Since the components of $\mathbf{G}$ are relative frequencies of rare alleles, it follows that their sum must equal one,

$$\mathbf{1}'\,\mathbf{G} = 1, \quad [9]$$

where $\mathbf{1}'$ is the vector $(1, 1, 1, \ldots . 1)$. We can easily modify Eq. 8 to incorporate this constraint, namely,

$$1 = \mathbf{1}'(1 - \mathbf{P}')^{-1}\mathbf{M}. \quad [10]$$

We are therefore able to determine the first element of $\mathbf{M}$, $2N\mu/K$, and thus estimate $g(j)$ from Eq. 8.

Estimation of K. In general, the sample size will be smaller than the population size and, thus, only the number of alleles in the sample (k) will be known. If the population is sampled without replacement, the probability of not detecting an allele that occurs as j copies in the population will be

$$g(j)\frac{\binom{2N-j}{2n}}{\binom{2N}{2n}} \quad [11]$$

where n is the sample size, and N is, as before, the size of the population. Thus, the probability of not detecting an allele in the sample that does occur in the population is

$$\sum_{j=1}^{2N} g(j)\frac{\binom{2N-j}{2N}}{\binom{2N}{2n}}. \quad [12]$$

Therefore, the expected number of alleles in the population is

$$E\,(k|K, 2N) = K\left\{1 - \sum_{j=1}^{2N} g(j)\frac{\binom{2N-j}{2n}}{\binom{2N}{2n}}\right\}. \quad [13]$$

By consideration of the method of moments type of estimation, a suitable estimator of K is therefore

$$\hat{K} = \frac{k}{1 - \sum_j \hat{g}(j)\,(1-f)^j} \quad [14]$$

where $f = n/N$. Estimation of μ follows easily from Eq. 10.

Extension of Model for Expanding or Contracting Populations. In the approach described above we have considered that population size is constant from generation to generation. It should be pointed out that this assumption is not necessary for the valid estimation of mutation rate using this approach. Apart from the assumption of the infinite alleles model, the only assumption necessary is that the distribution of relative frequency of rare alleles, $g(j)$, is constant from generation to generation. If the population has been changing in size at a constant rate for a number of generations, this assumption will still be reasonable. An estimate of mutation rate may then be obtained if the growth rate per generation is known. Eq. 2 can be modified to include the effect of an increasing population size as follows,

$$K_m g(1) = 2N_{m-1}\,\mu + K_m \sum_j g(j)P_{j1}, \quad [15]$$

where population size in the previous generation, N_{m-1}, can be calculated easily from a knowledge of the growth rate. This approach to the estimation of mutation rate can be considered to be an extension of that of Lea and Coulson (8).

APPLICATION OF MODEL

Data that are appropriate for calculation of mutation rate in this way must satisfy two main conditions: (*i*) the population must be defined and finite in size, (*ii*) there must be no immigration to the population. It is clear that acculturated populations, which have ill-defined population sizes and extensive immigration rates, do not meet these criteria. However, data from unacculturated tribal populations can be considered in many cases suitable for this approach. One of the best studied tribal populations is that of the South American Indian tribe, the Yanamamo, living in the area defined by the border between Venezuela and Brazil. This tribe has been the subject of a series of biological, medical, and sociological investigations by Neel and his group (9) over the past 17 years. Population size of the tribe is well defined, and immigration rates from either Western civilizations or from neighboring tribes have been judged to be almost nonexistent (10). We therefore use data from this tribe to illustrate the calculation of mutation rate. These data are described in more detail in the following paper by Neel and Rothman (7), which presents estimates of mutation rate using this and two other approaches, to data on 12 Amerindian tribes.

To calculate the mutation rate we must first estimate the transition probabilities, $\hat{P}_{ji}$, the distribution of the number of copies of rare alleles, $\hat{g}(j)$, the population size for a single generation, $\hat{N}$, and the number of alleles in the population, $\hat{K}$. The estimation of the distribution of the transition probabilities is probably the most involved. Normally, accurate determination of P_{ji} may be quite difficult, if not impossible, if the population of study has any appreciable structure. However, we are only concerned with the distribution of rare alleles in the population; under these circumstances we make use of the results of Li *et al.* (11), which show that the number of copies produced by each rare allele is distributed approximately geometrically. Thus,

$$A(x) = \begin{cases} 1 - \dfrac{b}{1-c}; j = 0 \\ bc^{j-1}; j \geq 1 \end{cases} \quad [16]$$

where b and c are parameters of the geometric distribution. For estimation of the parameters of the geometric distribution we use the data on family size distribution for the Yanamamo given by Neel and Weiss (10). Thompson and Neel (12) estimated the parameters of this geometric distribution, yielding values of $\hat{b} = 0.34$ and $\hat{c} = 0.40$ for the Yanamamo data. Practically identical estimates of b and c ($\hat{b} = 0.34$, $\hat{c} = 0.41$) were obtained (11) when account is taken of infanticide (12), a practice extant in the Yanamamo.

By Eq. 6 the transition probabilities can be calculated from the estimates of b and c to be

$$P_{ji} = \sum_{h=1}^{\min(i,j)} \binom{j}{h} \binom{i-1}{i-h} \left(1 - \frac{b}{1-c}\right)^{j-h} b^h c^{i-h}. \quad [17]$$

The expected relative frequency of rare alleles present as j copies in the population, $g(j)$, can be calculated from Eq. 8 and the constraint that $\sum_j g(j) = 1$. Table 1 shows this distribution for i, the number of copies, up to 30. The dimensions of the ($I - p'$) matrix were chosen empirically to be (30,30). The criterion for this choice was that our estimate of K remained constant to two significant figures over the range (25,25)–(30,30).

Population size in the Yanamamo has been estimated to be 15,000 (10). However, this figure includes individuals of both prereproductive age and postreproductive age and, therefore, includes individuals from more than one generation. Considering the reproductive span in the Yanamamo to be between the ages of 15 and 40, 48% lie within this cohort (10). The estimate of population is adjusted to reflect this, and is taken to be 7200.

Table 1. Estimated distribution of expected relative frequency of rare alleles with exactly i copies for the Yanamamo population

i	$g(i)$	i	$g(i)$	i	$g(i)$
1	0.425	11	0.016	21	0.004
2	0.130	12	0.013	22	0.003
3	0.087	13	0.012	23	0.003
4	0.063	14	0.010	24	0.003
5	0.048	15	0.009	25	0.002
6	0.038	16	0.008	26	0.002
7	0.031	17	0.007	27	0.002
8	0.026	18	0.006	28	0.001
9	0.021	19	0.005	29	0.001
10	0.018	20	0.005	30	0.001

Finally, estimation of K, the number of rare alleles present in the population, follows simply from Eq. 14. Twenty-eight loci were assayed, and the estimate of K obtained for the Yanamamo is 0.0995 rare alleles per locus.

Using the values of $\hat{N}$, $\hat{K}$, and $\hat{g}(j)$ we calculate the average mutation rate per generation per locus to be 0.145×10^{-5}. If we consider the Yanamamo as a growing population, it is more suitable to consider the extension of the model described earlier. Considering that the growth of this tribe is 0.75% per year (10), and assuming the generation time to be 25 years, our estimate of mutation rate becomes 0.175×10^{-5}. It can be seen that this value differs insignificantly from the estimate obtained assuming a constant population size. The Yanamamo yield the lowest mutation rate of all the 12 tribes considered in the following paper (7).

DISCUSSION

We have presented here an approach to the estimation of mutation rate that is particularly suitable for well-defined populations of finite size with no immigration. A number of assumptions must be made for this estimation procedure to be valid. (*i*) Each allele generated by mutation is uniquely identifiable. (*ii*) The distribution of the relative frequencies of rare alleles is constant from one generation to the next. (*iii*) The distribution of number of rare allele copies generated is known. For the Yanamamo we assume this distribution to be geometric. (*iv*) The rare alleles are neutral in selective value. (*v*) The sample is drawn at random from the population. Some of these are common to other estimation procedures. At this juncture it is appropriate to consider each of them in detail.

(*i*) **Infinite Alleles Model.** This assumption requires that each rare allele represents one mutational event. The most widespread biochemical technique used to identify mutant alleles is electrophoresis. Since an electrophoretic mobility class can represent a heterogeneous collection of alleles (see, e.g., refs. 13 and 14), it is possible that the rare alleles observed in the Yanamamo are polyphyletic in origin. However, we regard this as extremely unlikely. Two pieces of evidence support this conclusion. First, the data represent the total number of rare alleles seen for 26 different loci. From simple statistical considerations, the low number of alleles seen for this number of loci makes a polyphyletic origin extremely improbable. Second, the rare alleles found in Amerindian tribes often have a highly nonrandom distribution over the whole tribe and are usually restricted to a small geographical area encompassing small numbers of villages. An allele that is heterogeneous would not be expected to have such a nonrandom geographical distribu-

tion. Any bias introduced by our inability to detect a polyphyletic origin for rare alleles will tend to underestimate the true mutation rate, since the actual number of rare alleles will be larger than the observed number.

Although we can confidently assume that the rare alleles are homogeneous, we cannot make the same assumption for the common alleles. The technique of electrophoresis does not detect all genetic variation (13, 14) at a locus, and we expect that cryptic rare alleles are included in the common allele classes. Thus, only a proportion of the rare alleles will be detectable, and this will serve to underestimate the mutation rate. We regard this to be the most serious bias inherent to our model.

Recently it has become apparent that electrophoretic mobility at one locus may be influenced by variation at a second modifying locus. Evidence from *Drosophila* (V. Finnerty and G. Johnson, personal communication) suggests that modifying loci will increase the apparent number of electrophoretic variants for a locus. In general, this is also a potential source of bias in the estimation of mutation rates by the method described in this paper. At present we have no way of evaluating the importance of this possible effect.

(*ii*) **Constancy of Distribution of Allele Relative Frequencies.** For unacculturated tribal populations the equilibrium assumption is probably reasonable, as most evidence indicates that the structure and environment of the tribes have remained virtually unchanged for many generations. We have no way of critically evaluating the verity of this assumption, and so we consider here the possible biases introduced by the lack of an equilibrium. If the population has been growing or diminishing steadily in size, distribution of relative frequencies of rare alleles may still be constant in the population. The mutation rate may then be estimated without bias from Eq. **15**.

However, a bias in the mutation rate estimate will be introduced if the population has recently undergone a sudden increase or decrease in size. In such cases the matrix of P_{ij}s calculated from the offspring distribution will reflect only the rate of growth occurring in the present generation. For example, if the population has experienced a sudden increase in growth, the estimate of mutation rate will be biased downward. There is indication that the population size of the Yanamamo has recently increased (10), and this may account for the low mutation rate estimate in comparison to other South American tribes (7).

(*iii*) **Distribution of Rare Allele Copies.** The construction of the matrix of transition probabilities for the Yanamamo example relies on the assumption that the distribution of allele copies generated every generation is geometric. This assumption is justified theoretically and by the results from simulations (11) which show that the number of allele copies in the Yanamamo closely fits a geometric distribution. However, this fit was only approximate and so it is worthwhile to consider here the sensitivity of our mutation rate estimate to the characteristics of this distribution. Results (11) for the Yanamamo show that when the number of copies is large, the observed number of copies produced is consistently slightly higher than that predicted by a geometric distribution. Correction for this effect would increase the mutation rate. Alternatively, in other populations it may be that the distribution of allele copies approximates a Poisson distribution. In this case, the mutation rate would be overestimated using the geometric model.

(*iv*) **Neutrality of Mutant Alleles.** The distribution of the relative frequencies of rare alleles must be estimated from the offspring distribution on the population. This approach assumes that the behavior of all alleles is equal and neutral. Such an assumption is common to the earlier approach for estimating mutation rate (2). Some of the rare mutant alleles are probably selected against. If these are recessive to the normal allele in selective effect, as is likely, their selective disadvantages will never be manifested, as their rarity determines that they will always occur with a normal allele in a heterozygote. Thus, most of the mutant alleles can, for practical purposes, be considered to be neutral in effect and to act independently of each other. Any manifestation of selection against the mutant alleles will render our estimate of mutation rate an underestimate, since the mutation rate required to maintain a set of deleterious alleles in the population will be higher than that required for a set of neutral alleles given the same frequency distribution.

(*v*) **Randomness of Sample.** It is clear that in this approach to the estimation of mutation rate, as in so many other statistical procedures, the sample must be drawn at random from the population. However, it is equally clear that this requirement is rarely, if ever, met in sampling tribal populations. For the Yanamamo data, used here as an example, the sample is definitely nonrandom. This tribe was sampled by villages, and in any one village, villagers tended to be sampled in family groups (9). The effect of this nonrandomness is to underestimate the mutation rate. From a consideration of Eq. **14**, it can be seen that the estimate of the number of alleles in the population, $\hat{K}$, is a function of f, the ratio of sample size to population size. The true random sample is smaller than the actual sample. Consequently, the estimate of the number of alleles in the population is underestimated and this, in turn, will lead to an underestimate of the mutation rate.

It is appropriate to mention here a further effect of the sampling procedure that also serves to underestimate mutation rate. The mutation rate estimated for the Yanamamo is an average over 28 loci. However, the sample size for each locus varied greatly. This variation was in part related to the difficulty of assaying some enzymes and sensitivity to storage between the time the sample was taken and its arrival in the laboratory. For example, the sample size for the 2,3-diphosphoglycerate mutase locus was 149, while sample size for the albumin locus was 3504. In our calculations we have used the arithmetic mean sample size for all loci. Although this simplification is necessary to render the calculations tractable, it also biases the estimate of mutation rate downwards. Eq. **14** shows that $\hat{K}$ is a nonlinear function of sample size. As sample size is increased, the rate of increase of the estimate of the number of alleles in the population ($\hat{K}$) is reduced. Thus, in any average estimate, unusually low sample sizes will contribute disproportionately to $\hat{K}$. The use of the average sample size to calculate $\hat{K}$ reduces the contribution of the small sample sizes of particular loci and therefore biases the estimates of $\hat{K}$ and mutation rate downward.

Conclusions

We have described here an approach for estimation of mutation rate which is based on a determination of the number of alleles in a population sample and on the distribution of the number of offspring in the population. Both statistics are comparatively easy to obtain. The estimate of mutation rate obtained by this approach is subject to a number of biases. However, most of the biases described above have the effect of underestimating the mutation rate.

We thank J. V. Neel and P. Smouse for helpful comments on the manuscript. This work was supported in part by Department of Energy contract EY-77-C-02-2828.

1. Cavalli-Sforza, L. L. & Bodmer, W. R. (1971) *The Genetics of Human Populations* (W. H. Freeman, San Francisco).
2. Kimura, M. & Ohta, T. (1969) *Genetics* **63**, 701–709.
3. Neel, J.V. (1973) *Proc. Natl. Acad. Sci. USA* **70**, 3311–3315.

4. Nei, M. (1977) *Am. J. Human Genet.* **29,** 225–232.
5. Wright, S. (1931) *Genetics* **16,** 97–159.
6. Ewens, W. J. (1972) *Theor. Popul. Biol.* **3,** 87–113.
7. Neel, J. V. & Rothman, E. D. (1978) *Proc. Natl. Acad. Sci. USA* **75,** in press.
8. Lea, D. E. & Coulson, C. A. (1949) *J. Genet.* **49,** 264–285.
9. Neel, J. V. (1978) *Annu. Rev. Genet.* **12,** in press.
10. Neel, J. V. & Weiss, K. M. (1975) *Am. J. Phys. Anthropol.* **42,** 25–52.
11. Li, F. H. F., Neel, J. V. & Rothman, E. D. (1978) *Am. Nat.* **112,** 83–96.
12. Thompson, E. A. & Neel, J. V. (1978) *Proc. Natl. Acad. Sci. USA* **75,** 1442–1445.
13. Coyne, J. (1976) *Genetics* **84,** 593–604.
14. Singh, R. C., Lewontin, R. C. & Felton, A. A. (1976) *Genetics* **84,** 607–629.