*Proc. Natl. Acad. Sci. USA*
Vol. 75, No. 3, pp. 1442–1445, March 1978
Genetics

# Probability of founder effect in a tribal population

(Amerindian/genetic variant/branching processes/private variants/demographic history)

E. A. THOMPSON† AND J. V. NEEL

Department of Human Genetics, University of Michigan Medical School, 1137 E. Catherine Street, Ann Arbor, Michigan 48109

*Contributed by James V. Neel, December 8, 1977*

**ABSTRACT** **When an unusually high frequency of an allele is encountered in a population, "founder effect" is often invoked as an explanation. As usually used, the term implies the disproportionate increase through chance (rather than selection) of an allele contributed to the population by a particular ancestor. While genetic theory leaves no doubt this is a possible explanation, problems arise when we try to determine how likely this explanation is for any specific finding in any specific, finite population, i.e., just how rare is this rare event? In this communication we consider the question in the context of Amerindian tribal populations, deriving specific probabilities under defined conditions. Our interest in the question has been whetted by the finding to date of some eight possible examples of a founder effect in studies of twelve different tribes.**

## Branching process model for rare variants

A number of genetic variants whose distributions are restricted to single tribes have been found in Amerindian populations (1). Some of these have attained allele frequencies greater than 0.01, i.e., polymorphic frequencies, and thereby raise the question of what demographic and genetic forces can account for this finding. Although most models for allelic variability in populations consider allele frequencies in a population of constant size, for rare variant alleles a branching process model for the number of replicate copies seems more appropriate. Such models have been considered by, amongst others, Fisher (2) and Karlin and MacGregor (3). Here we consider an extension of the model used by Thompson (4, 5) in estimating the age and rate of increase of a rare variant allele.

We shall assume that at generation $j$ the number of replicate copies, $k$, produced independently by any given variant gene in the current generation has probability distribution

$$p^{(j)}(k) = \begin{cases} r_j & k = 0 \\ (1 - r_j)(1 - c_j)c_j{}^{k-1} & k = 1, 2, \ldots \end{cases} \quad [1]$$

This is the two-parameter geometric distribution parameterized as in Keiding and Nielsen (6). The parameter $r_j$ is the probability of immediate extinction of any given gene, and $c_j$ is the geometric parameter. The mean number of replicates produced, $m_j$, is given by

$$m_j = (1 - r_j)/(1 - c_j), \quad [2]$$

the mean conditional on a non-zero number of replicates, $w_j$, is

$$w_j = 1/(1 - c_j), \quad [3]$$

and the variance $v_j$ is given by

$$v_j + m_j{}^2 = m_j \frac{(1 + c_j)}{(1 - c_j)} = m_j(2w_j - 1). \quad [4]$$

Under single-generation distributions of the form [1], Keiding and Nielsen (6) note that the cumulative distribution for the number of replicate copies after $t$ generations is also of the two-parameter-geometric form, and show that the relationship between the $t$-generation parameters $R_t$, $G_t$, $M_t$, and $W_t$ and their single-generation equivalents $r_j$, $c_j$, $m_j$, and $w_j$ are given by

$$M_t = \prod_1^t m_j,$$

$$\frac{1}{1 - R_t} = 1 + \sum_{j=1}^{t} \frac{r_j}{(1 - r_j)M_{j-1}}, \text{ and} \quad [5]$$

$$W_t = \frac{M_t}{(1 - R_t)} = \frac{1}{(1 - G_t)}.$$

In the special case $m_j = m$ and $c_j = c$ for all $j$, the formulae **5** reduce to those of Thompson (4):

$$M_t = m^t, R_t = \frac{(m^t - 1)(1 - (m - 1)(1 - c)/c)}{((m - 1)(1 - c)/c) + (m^t - 1))}. \quad [6]$$

We shall be interested in the possibility of variants becoming replicated in large numbers. At $t$ generations the probability of $k$ replicates is

$$P_t(k) = (1 - R_t)(1 - G_t)G_t{}^{k-1} \quad k = 1, 2, \ldots$$

and the probability of more than $k$ replicates is thus

$$Q_t(k) = \sum_{i=k+1}^{\infty} P_t(i) = (1 - R_t)G_t{}^k = (1 - R_t)\left(1 - \frac{1}{W_t}\right)^k. \quad [7]$$

In a series of studies, our associates and ourselves have documented eight instances in which the members of a tribe of South American indians, or of several closely related tribes, possess apparently unique alleles in frequencies greater than 0.01, i.e., "private" genetic polymorphisms (1). The variant alleles responsible for these polymorphisms must necessarily be old, and over long periods of time a tribe will not have enjoyed a constant rate of growth. An important aspect of the problem is thus of the effect of fluctuations in the parameter $m_j$ on the survival probability of a variant allele. The offspring distribution has two independent parameters, most conveniently taken as $m_j$ and $c_j$; the parameter $c_j$ defines the offspring distribution conditional on a non-zero number of replicates. We have therefore considered the effect of varying the growth rate $m_j$ over a population's history, subject to a constant value of $c_j$. The probability of immediate extinction $r_j$ and the second moment $(v_j + m_j{}^2)$ therefore change linearly with changes in $m_j$ (Eqs. **2** and **4**).

In practice we estimate the current number of replicates of a variant allele in the adult generation. The parameters therefore relate to the distribution of gene copies in the adult generation resulting from a single gene in an adult of the previous generation. The variants are of course carried in diploid indi-

† Present address: King's College, Cambridge CB2 1ST, England.

Table 1. Simulation and expected fates of 280 genes whose possessors were aged initially 10–19 years

| No. of copies | Simulation | | Expected ($c$ = 0.40) | | | |
|---|---|---|---|---|---|---|
| | Run 1 | Run 2 | $m$ = 1 | $m$ = 1.05 | $m$ = 1.1 | $m$ = 1.15 |
| 0 | 247 | 231 | 256.0 | 243.6 | 228.5 | 211.3 |
| 1–50 | 28 | 40 | 23.7 | 34.6 | 44.8 | 50.5 |
| 51–100 | 4 | 9 | 0.3 | 1.7 | 5.7 | 13.4 |
| >100 | 1 | 0 | — | 0.1 | 1.0 | 4.8 |

viduals, and a basic assumption of a branching process model is that the allele frequency is sufficiently low for there to be few homozygotes. In this case the parameters for a geometric distribution of diploid offspring may be related to those for variant replicates, and hence family size data may be used to estimate the required parameters. Thompson (4) estimated $c = 0.40$ from data on family size.

The geometric offspring distribution has the convenient property that the form is unaltered by incorporating a phase of random survival. If variant alleles survive from birth to adulthood with probability $q$, and the distribution of replicates born to each adult is geometric with parameters $m$ and $c$, then the birth-to-birth distribution is geometric with parameters $m' = qm$ and $c' = c$ and the adult-to-adult distribution has $m'' = qm$ and $c'' = cq/(1 - c + cq)$. Thus,

$$m' = m'' \text{ and } \frac{c''}{1-c''} = q \cdot \frac{c'}{(1-c')}. \qquad [8]$$

New variants arise, of course, in newborn individuals, but must survive to adulthood in the initial generation if they are to become replicated in large numbers. The ratio of new variants arising to those reaching adulthood is $1:q$, and the counts of variant replicates at birth and in adults in the current generation must be in the same ratio.

## Some results from simulation

In addition to the derivation of founder-effect probabilities on the basis of the above mathematical model, the question has also been approached through computer simulation of an Amerindian tribe. The details of the simulation have been described by Li *et al.* (7). The simulation, modeled after the observed demographic parameters of an Amerindian tribe studied in some detail, the Yanomama, allows us to follow the fate of the population for some 400 years. Each member of the founding population of 451 persons, distributed among four villages, is assigned at the outset four pairs of genes uniquely identified by numbers, each pair consisting of one odd- and one even-numbered allele. All alleles were assigned the same survival value.

First we consider the fate over the 16 generations encompassed by the simulation of alleles carried by subadults and young adults. These are defined as individuals aged 10–19 years; there are 70 such individuals in the initial population. They have survived the relatively high mortality of infancy and childhood and are now in a period of relatively low mortality (8). We ask how many of the odd-numbered alleles present in each of these individuals survive the 16 generations and in what numbers are the survivors represented? The results are shown in Table 1. Because the fates of the four odd-numbered alleles of a single individual are not independent, the 280 entries of Table 1 cannot be considered as independent trials.

Table 2. Probability of a rare variant exceeding the given number of replicates at the given generation*

| $m$ | $t$ | No. of copies >0 | >100 | >400 | >1000 |
|---|---|---|---|---|---|
| | | Simulation $c$ = 0.67 | | | |
| 0.98 | 16 | 0.0242 | $0.0^372$ | $0.0^719$ | $0.0^{16}13$ |
| | 100 | 0.0013 | $0.0^339$ | $0.0^411$ | $0.0^880$ |
| | 400 | $0.0^514$ | $0.0^648$ | $0.0^720$ | $0.0^{10}32$ |
| 1.0 | 16 | 0.0296 | $0.0^215$ | $0.0^618$ | $0.0^{14}27$ |
| | 100 | 0.0049 | $0.0^230$ | $0.0^369$ | $0.0^438$ |
| | 400 | 0.0012 | $0.0^211$ | $0.0^375$ | $0.0^336$ |
| 1.05 | 16 | 0.0441 | $0.0^257$ | $0.0^413$ | $0.0^{10}61$ |
| | 100 | 0.0246 | 0.0241 | 0.0228 | 0.0204 |
| | 400 | 0.0244 | 0.0244 | 0.0244 | 0.0244 |
| 1.1 | 16 | 0.0615 | 0.0160 | $0.0^328$ | $0.0^787$ |
| | 100 | 0.0487 | 0.0487 | 0.0487 | 0.0486 |
| | 400 | 0.0487 | 0.0487 | 0.0487 | 0.0487 |
| | | Adult-to-adult $c$ = 0.40 | | | |
| 1.0 | 16 | 0.0857 | $0.0^411$ | $0.0^{16}23$ | 0.0 |
| | 100 | 0.0148 | $0.0^233$ | $0.0^438$ | $0.0^851$ |
| | 400 | 0.0037 | $0.0^226$ | $0.0^384$ | $0.0^488$ |
| 1.05 | 16 | 0.1302 | $0.0^328$ | $0.0^{11}27$ | $0.0^{27}26$ |
| | 100 | 0.0755 | 0.0713 | 0.0600 | 0.0425 |
| | 400 | 0.0750 | 0.0750 | 0.0750 | 0.0750 |
| 1.1 | 16 | 0.1840 | $0.0^231$ | $0.0^715$ | $0.0^{18}32$ |
| | 100 | 0.1500 | 0.1498 | 0.1494 | 0.1484 |
| | 400 | 0.1500 | 0.1500 | 0.1500 | 0.1500 |

* Computed from Eqs. 6 and 7.

It is of interest to contrast this "observation" with theoretical prediction based on Eq. 6 for the case of $c = 0.40$ and constant $m$. The comparison is not entirely appropriate since the theoretical formulation assumes a cohort of individuals immediately prior to reproduction, whereas some members of the simulation cohort will die before that time. However, as shown in Table 1, the agreement is good at $m = 1.1$ and 1.05. These values of $m$ are consistent with recent Yanomama history (8), but it is impossible to state how long such values have obtained.

A second question is of the fate in the simulation population of a mutant allele introduced into a newborn infant. There were 177 instances in which among the offspring born to a couple in the first generation, a single copy of a given gene was transmitted. This was considered equivalent to the introduction of a mutant into a newborn child. Each of these alleles was then followed until its extinction or the completion of the 400-year run. Maximum likelihood estimation of the parameters of the offspring distribution, on the assumption that these were constant, yielded $c = 0.6723$ and $m = 0.9779$.

At the completion of generation 16 there were four surviving mutants, none represented by more than 50 copies. Prediction from Eq. 6 with the above $c$ and $m$ was of survival of a proportion 0.0226 of such mutants, none (to 9 decimal places) to exceed 50 copies in number. The agreement between observation and prediction is satisfactory, and we have therefore proceeded to consider the probabilities generated by the mathematical formulation over much longer periods of time and therefore for larger numbers of replicate copies (Table 2).

We note that an adult-to-adult $c$ of 0.40 and a birth-to-birth value of 0.67 implies, from Eq. 8, a birth-to-adult survival, $q$, of approximately 1/3. Neel and Weiss (8) give values of 0.34 and 0.36 for female and male survival probabilities, respectively, to the midreproductive period.

### Demographic expansion and founder effect

At certain times in its history, for example, following its entry into new territory, an Indian tribe may be expected to enjoy a period of rapid expansion. This may often be followed by a long period of relatively constant size. Consider a single gene at the beginning of the expansion period of $t$ generations, which is to be characterized by cumulative parameters $M$, $R$, and $W = M/(1-R)$ in the notation of the first section, and suppose the second period gives rise to parameters $M^*$, $R^*$, and $W^*$. If $g(Z)$ is the generating function for numbers of replicates produced from a single initial gene over the first period, then $g(Z) = R + (1-R)Z/(W-(W-1)Z)$. The overall generating function over both periods is $g(g^*(Z))$, in which $g^*(Z) = R^* + (1-R^*)Z/(W^*-(W^*-1)Z)$ is the generating function over the second period.

Hence, the overall mean is given by

$$M^{**} = \frac{\delta}{\delta Z}[g(g^*(Z))]|_{Z=1} = MM^*$$

and the overall extinction probability is given by

$$R^{**} = g(g^*(0)) = g(R^*) = R + \frac{(1-R)R^*}{(W-(W-1)R^*)}.$$

If $D = 1-R$ and $D^* = 1-R^*$ are survival probabilities over the two periods, the net survival is given by

$$D^{**} = 1 - R^{**} = D\left\{1 - \frac{(1-D^*)}{(W-(W-1)(1-D^*))}\right\}. \quad [9]$$

For a period of rapid population expansion followed by a long period of approximately constant size,

$$M^* \approx 1,\ D^* \approx 0, \text{ and } D^{**} \approx DWD^* = MD^*. \quad [10]$$

That is, the survival probability is increased precisely to the extent of the initial population expansion. Although a period of rapid expansion will thus considerably enhance the survival probability of any variant arising at its commencement, and may be the explanation of many of the observed cases of founder effect, we shall see that such an expansion has little effect on the total number of variants we expect to see replicated in large numbers (see ref. 9).

Rather than, or in addition to, a single period of expansion, a plausible model for the long-term history of a tribe is of a gradual expansion punctuated by periodic sharp reverses (epidemics or famine). We shall assume that between reverses the natural rate of population increase is $m$ ($\geq 1$) and that at every $L$ generations there is a crash, providing an overall expectation of constant population size. Thus, if generation $j$ is not a "disaster" generation, we have $m_j = m$, $c_j = c$, and $r_j = r$, in which $(1-r) = m(1-c)$, but when it does experience a precipitous population decline $m_j = (1/m)^{L-1}$, $c_j = c$, and thus $r_j = 1 - (1-r)/m^L$.

We consider a variant that arose $t$ generations ago, $g_1$ generations before a population reverse, and assume the state of the population is now $g_2$ generations since a reverse. Thus, $t = g_1 + (f-1)L + g_2$ for some positive integer $f$. Then substituting in Eq. 5 we obtain

$$M_t = m^{g_1+g_2-L} \text{ and}$$

$$1 - R_t = 1 + \frac{f}{(1-r)}(m^L - (1-r)) + \frac{r}{(m-1)(1-r)m^{g_1-2}}\{(f-1)(m^{L-1}-1) + m^{g_1-1} - 1 + m^{L-1} - m^{L-1-g_2}\}. \quad [11]$$

Table 3. Number of generations before the next reverse, $g_1$, at which a variant must arise in a cycle length $L$, to have at least the same survival probability as in a constant ($m = 1$) population ($c = 0.40$)

| $m$ | $L = 10$ | $L = 20$ | $L = 40$ |
|---|---|---|---|
| 1.01 | 6 | 11 | 22 |
| 1.02 | 6 | 11 | 22 |
| 1.05 | 6 | 12 | 24 |
| 1.08 | 6 | 12 | 26 |
| 1.10 | 6 | 13 | 27 |
| 1.15 | 6 | 13 | 29 |

If the variant has undergone an exact number of cycles, then $g_1 + g_2 = L$, $M_t = 1$, and

$$\frac{1}{(1-R_t)} = 1 + \frac{f(m^L-1)(r+m-1)}{(1-r)(m-1)m^{g_1-1}}.$$

Hence

$$W_t = \frac{M_t}{(1-R_t)} = \frac{1}{(1-R_t)} = 1 + \frac{fc(m^L-1)}{(1-c)(m-1)m^{g_1-1}}. \quad [12]$$

Recalling that $R_t$ is the probability of extinction of the variant before age $t$ and $W_t$ is the expected number of replicates conditional on nonextinction, we see that the effect of a variant arising early in a cycle is to increase survival probability but decrease expected numbers conditional on survival according to the factor $m^{g_1-1}$ in Eq. 12. The effect of cycles generally is to decrease survival probability and to increase expected numbers, except for those variants arising early in the cycle, since for given $L$, $(m^L-1)/(m-1)$ is an increasing function of $m$. Table 3 shows at what points in the cycle a variant must arise to have the same survival probability as in a constant ($m = 1$) population. We note also that the expected numbers generated by Eq. 12 are linear in $f$, the number of cycles elapsed.

As described in the section *Some results from simulation*, Table 2 gives the probabilities of a variant being present in a population with more than a given number of copies at a given generation, assuming a constant value of $m$. The way in which these values would be modified by a superimposed cyclic pattern has been discussed, but we should consider also the actual numerical values of these probabilities. Table 2 gives the probabilities at three epochs. The first (16 generations) corresponds to the simulation. The last (400 generations) is a period of approximately 10,000 years, which we take to be an upper bound on the age of a variant arising since the differentiation of the current South American tribes. One hundred generations provides a convenient intermediate point. With respect to the number of copies, more than 0 copies is simply survival, whereas more than 100 would correspond in most tribes to a well-established private polymorphism. A number in excess of 1000 is an attribute of presumably very old polymorphisms, thus far encountered only twice in Amerindians, and must be an extreme value for any variant restricted to a single tribe.

Two values of $c$ have been considered, one corresponding to the birth-to-birth distribution ($c = 0.67$ from simulation) and one to the adult-to-adult distribution ($c = 0.40$), the difference being given by a birth-to-adult survival rate of approximately $\frac{1}{3}$. Table 2 shows that, as expected, survival of a mutant is greatly enhanced by its introduction into an adult migrating into the population, rather than into a newborn, but that it requires a longer period before the number of replicates in adults

reaches large numbers, both effects resulting from the high prereproductive mortality.

Theoretically there is a marked difference between the case of an expanding ($m > 1$) and a stable or declining population ($m \leq 1$). In the former case there is a non-zero probability of ultimate survival, and probabilities of exceeding any given value increase to this limit. It is not possible, of course, that Amerindian tribes could increase at a rate of $m = 1.1$ or even 1.05 for 400 generations (these values are given only for completeness), but we see that even in 16 generations 1.5% of new mutants will exceed 100 copies if $m = 1.1$. However, the probability within this period of large numbers of replicates is small. An average $m$ value of 1.0 could, of course, be maintained over long periods. In this case the population (and hence all variants) must eventually become extinct, and the probability of exceeding any specified number of copies will attain a maximum before decreasing to zero. (This is an inevitable feature of a branching process model.) Note that at 100 generations 3 in 1000 variants will exceed 100 copies, while at 400 generations only 1 in 1000 will do so, but in the latter case over $\frac{1}{3}$ of those exceeding 100 copies also exceed 1000, while in the former only 1 in 79 does so.

## Conclusions

In tribal populations with relatively high prereproductive mortalities, the probability that any specific variant will ever attain very large numbers of replicates is small, though in a population expanding rapidly over a short period some variants may relatively quickly attain polymorphic frequencies. In such a population a variant introduced by a migrating adult has a much higher survival probability than a mutation arising in a newborn, but if the tally is from adult-to-adult a correspondingly longer period is required before a given number of replicates can be attained. A cyclic pattern of population increase can also considerably enhance survival probabilities for mutants arising at certain points in the cycle, and the demographic history of a population is thus an important factor in assessing the probability that variants will become replicated in large numbers.

We thank Mrs. Betty Y. Hsiao for outstanding computational assistance. This research was supported by Energy Research and Development Administration Contract EY-77-C-02-2828 and National Science Foundation Grant BMS-74-11823 and was carried out while E.A.T. was a Visiting Postdoctoral Scholar, Department of Human Genetics, University of Michigan.

1. Neel, J. V. (1977) *Am. J. Hum. Genet.*, in press.
2. Fisher, R. A. (1930) *The Genetical Theory of Natural Selection* (Clarendon Press, Oxford).
3. Karlin, S. & MacGregor, J. (1967) in *Proceedings of the Fifth Berkeley Symposium on Statistics and Probability*, eds. leCam, L. M. & Neyman, J. (University of California Press, Berkeley, CA), pp. 415–438.
4. Thompson, E. A. (1976) *Am. J. Hum. Genet.* **28**, 442–452.
5. Thompson, E. A. (1977) in *Measuring Selection in Natural Populations*, eds. Christiansen, F. B. & Frenchel, T. M., Lecture Notes in Biomathematics (Springer-Verlag, Berlin), Vol. 19, pp. 531–544.
6. Keiding, N. & Nielsen, J. E. (1975) *J. Appl. Prob.* **12**, 135–141.
7. Li, F., Neel, J. V. & Rothman, E. D. (1978) *Am. Nat.*, in press.
8. Neel, J. V. & Weiss, K. M. (1975) *Am. J. Phys. Anthropol.* **42**, 25–52.
9. Neel, J. V. & Thompson, E. A. (1978) *Proc. Natl. Acad. Sci. USA* **75**, in press.