Proc. Natl. Acad. Sci. USA
Vol. 75, No. 4, pp. 1904–1908, April 1978
Genetics

Founder effect and number of private polymorphisms observed in Amerindian tribes

(population genetics/blood serum protein/erythrocyte enzymes/electrophoresis)

JAMES V. NEEL AND E. A. THOMPSON*

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 **In studies extending over the past dozen years, we have observed eight examples of "private" genetic polymorphisms in 12 Amerindian tribes surveyed for electrophoretic variants of an average of 25 proteins. Each of these is presumed to trace to a single mutation. In a preceding communication [Thompson, E. A. & Neel, J. V. (1978) *Proc. Natl. Acad. Sci. USA* 75, 1442–1445] the statistical theory was developed for estimating the likelihood of such a founder effect in a tribal population of this type. In this paper that theory is applied to the distribution defined by these eight variants. It is demonstrated that on the assumption that the phenotypes in question are selectively neutral, such findings are most compatible with a mutation rate of 7×10^{-6}/locus per generation. This figure applies only to variants that can be detected by the electrophoretic technique.**

In a series of recent studies of blood samples obtained from members of relatively intact and unmixed Amerindian tribes in Central and South America, we have encountered a number of "private" genetic polymorphisms. These are defined as apparently unique genetic variants that, within a single tribe or several adjacent and related tribes, attain allele frequencies greater than 1%. These polymorphisms, all of blood serum proteins or erythrocyte enzymes, have been recognized on the basis of their kinetic or electrophoretic characteristics. At present, the precision of the detection of these variants is much greater with the electrophoretic than with the other techniques; this fact permits certain quantitative treatments of the electrophoretic variants. In this communication we restrict ourselves to a consideration of the implications of the eight traits identified by atypical electrophoretic behavior. These eight traits have been observed in a survey of 12 tribes for electrophoretic variants of an average of 25 proteins.

The frequency with which private polymorphisms such as these are encountered in a defined population is a function of mutation rate, the type of selection operating on the phenotype in question, and the size of the population in which the trait has arisen. The latter cannot be estimated with the necessary accuracy for recently conglomerate, civilized populations, but for tribal populations, to a first approximation we may consider the tribe as the breeding unit. Estimates of tribal size are available for all the tribes we have studied. In the companion paper to this communication (1), a mathematical treatment has been developed of the probability of survival of a mutant gene in a population with the breeding structure of an Amerindian tribe, under a variety of assumptions regarding the growth characteristics of the population. We therefore are now in the position to ask the following question: What combination of mutation rates, selective pressures, and population growth can account for these findings in Amerindians? In a subsequent paper we will estimate the average mutation rate per locus per generation for characteristics of this type in these Amerindian populations, and thereby close in on the question of the nature of the selection, if any, to which these traits are subject (J. V. Neel, E. D. Rothman, and J. Adams, unpublished data).

THE DATA

The 12 tribes that have been surveyed are the Ayoreo, Baniwa, Cayapo, Guaymi, Kraho, Macushi, Makiritare, Central Panoa, Piaroa, Wapishana, Xavante, and Yanomama. The serum proteins and erythrocyte enzymes that have been examined for electrophoretic variants are the following: *Erythrocyte Proteins:* acid phosphatase-1 (ACP_1), adenosine deaminase (ADA), adenylate kinase-1 (AK_1), carbonic anhydrase-1 (CA_1), carbonic anhydrase-2 (CA_2), 2,3-diphosphoglycerate (DPG), esterases A_1, A_2, and A_3 ($ESA_{1,2,3}$), esterase D (ESD), galactose-1-phosphate uridyltransferase (GALT), hemoglobin A (Hbα and Hbβ), hemoglobin A_2 (Hbα and Hbδ), isocitrate dehydrogenase (ICD_s), lactate dehydrogenase (LDH_A and LDH_B), malate dehydrogenase (MDH_2), nucleoside phosphorylase (NP), peptidase A (PEPA), peptidase B (PEPB), phosphoglucomutase-1 (PGM_1), phosphoglucomutase-2 (PGM_2), phosphogluconate dehydrogenase (PGD), phosphohexose isomerase (PHI), and triosephosphate isomerase (TPI). *Serum Proteins:* albumin (ALB), ceruloplasmin (CP), haptoglobin (HP), and transferrin (TF). The products associated with at least 28 genetic loci are presented. Not all tribes were surveyed for all proteins. A complete review of the findings is in press (2).

For the mathematical treatment that follows we require an estimate of the number of copies of the mutant allele in the adult generation of the tribe in which each of the alleles in question was encountered. Table 1 lists the eight traits under consideration, using the nomenclature of the preceding paragraph, the tribe or tribes in which each occurs, and the estimated *present* size of the tribe in question. The latter is based on an extensive personal field experience as well as on surveys by government and mission groups and conversations in the field with knowledgeable persons, and is a composite opinion. Even so, these estimates represent approximations. From detailed studies of one tribe, the Yanomama, we estimate that the proportion in the adult generation, defined as all individuals between the ages of 15 and 40 inclusively, is approximately 48% of the total population (3). On this basis, the number in the adult generation (N) has been estimated for each tribe in which a private polymorphism has been encountered. An allele frequency has been estimated from the number of hetero- and homozygotes observed and total sample size; from this and N we estimate the total number of adult-generation copies of the allele in question. In two instances, the allele in question occurs in two tribes. In the case of $ESA_1{}^{D\ MAC\ 1}$, this is due to the recent admixture of these two tribes. In the case of $CP^{A\ CAY\ 1}$, the

* Present address: King's College, Cambridge, England.

Table 1. Estimated number of alleles for each of the electrophoretically defined private polymorphisms encountered in 12 tribes tested on average for 25 systems

Allele	Tribe	Tribal size	N^*	Sample size	Affected individuals	$\hat{P}$	No. copies in N	
$ALB^{YAN\ 2}$	Yanomama	15,000	7,200	3,504	461 YAN 2/+ 30 YAN 2/YAN 2	0.074	1,066	
ALB^{MAKU}	Wapishana	2,000	960	623	22 MAKU/+ 1 MAKU/MAKU	0.019	37	
$CAII^{BAN\ 1}$	Baniwa	1,500	720	377	37 BAN 1/+ 2 BAN 1/BAN 1	0.054	78	
$CP^{A\ CAY\ 1}$	Cayapo	1,500	720	668	67 A CAY 1/+	0.050	72	85
	Xavante	1,700	816	457	7 A CAY 1/+	0.008	13	
$ESA_1{}^{D\ MAC\ 1}$	Macushi	4,000	1,920	498	47 D MAC 1/+ 1 D MAC 1/D MAC 1	0.049	188	234
	Wapishana	2,000	960	614	27 D MAC 1/+ 1 D MAC 1/D MAC 1	0.024	46	
$LDH_B{}^{GUA\ 1}$	Guaymi	30,000	14,400	484	61 GUA 1/+ 4 GUA 1/GUA 1	0.071	2,045	
$PEPA\ 2^{WAP\ 1}$	Wapishana	2,000	960	614	17 2 WAP 1/+ 3 2 WAP 1	0.016	31	
$PEPB^{PAN\ 1}$	Central Pano	18,000	8,640	335	16 PAN 1/+	0.024	415	

References to detailed descriptions of each of these variants are to be found in ref. 1.
* Number in adult generation.

tribes in question are adjacent and of the same language family; we assume the mutation occurred prior to the separation of the two tribes.

THE THEORY

In a previous paper (1) we considered the probabilities of any given mutant allele surviving to become replicated in large numbers. Here we consider the total number of such variants we should expect to see in a collection of populations. In the previous paper we derived the probability that within a single population a variant arising t generations ago will now have more than k replicates;

$$Q_t(k) = (1 - R_t)\left(1 - \frac{1}{W_t}\right)^k \qquad [1]$$

in which R_t is the probability of extinction over the t generations and W_t is the expected number of replicates at the end of that period conditional upon nonextinction. The expected number of replicates is $M_t = W_t\,(1 - R_t)$. The input of new variants into the population is independent of the replication process; in any generation, each gene has a probability μ of being a new mutation.

Hence, conditional on a population size N_t (genes) t generations ago, the number of new mutants is binomially distributed with index N_t and mean $N_t\mu$, where μ is the mutation rate. The number now replicated in numbers greater than k is binomially distributed with index N_t and mean $N_t\mu Q_t(k)$, or since μ and $Q_t(k)$ are both small, approximately Poisson with this same mean. Summing over generations, we obtain, conditional upon values of N_t, a Poisson distribution with mean $\{\Sigma N_t Q_t(k)\}\mu$. The successive values of N_t are presumably also subject to stochastic variation, and moreover are correlated. The mean for numbers greater than k is thus

$$\mu \sum_t E(N_t)Q_t(k) \qquad [2]$$

and the variance is of the same order, though inflated by positive correlations between successive N_t values. Fluctuations of the total population will be proportionately smaller than stochastic variation in the number of replicates of a given rare variant, and to a first approximation expression [2] may also be taken as the variance of the number of variants with more than k replicates.

Now in the absence of selection $N_t \approx N/M_t$, where N is the total number of genes in the *current* population, and expression [2] reduces to

$$\mu N \sum_t (1 - R_t)\left(1 - \frac{1}{W_t}\right)^k \Big/ M_t = \mu N \sum_t \left\{\left(1 - \frac{1}{W_t}\right)^k \cdot \frac{1}{W_t}\right\}. \qquad [3]$$

This sum is in general unbounded, but we are interested in *private* variants: variants present in a single tribe or closely related tribes. These have presumably arisen subsequent to tribal differentiation. We therefore compute the truncated sum

$$S = \sum_{t=0}^{T} \frac{1}{W_t}\left(1 - \frac{1}{W_t}\right)^k. \qquad [4]$$

This is the contribution, as a proportion of $N\mu$, to a class of variants now with more than k replicates from those variants age less than T in a single population. Since both means and variances sum over independent populations, it is equally the contribution to the total number of such variants in a group of populations, undergoing the same average demographic pattern, where N now refers to the total number of current genes in all populations. In Fig. 1, S is considered as a function of T. For small k the largest contributions will come from the recent generations, where a larger proportion of variants survive. For large k only the old variants will contribute, the recent ones not yet having achieved large numbers. W_t is monotonic increasing in t, and the maximal contribution comes from the generation in which $W_t = k + 1$.

In the previous paper we considered the effect of a period of rapid population expansion and of patterns of cyclic growth on the survival probability of a variant. In the figures and tables to follow, we again consider the results of these same demographic patterns. In Fig. 1 the consequences of a rapid expansion of 20% per generation for 20 generations, followed by $m = 1$ for 180 generations, is compared with the consequences of an equivalent gradual increase of 2% for 200 generations. In

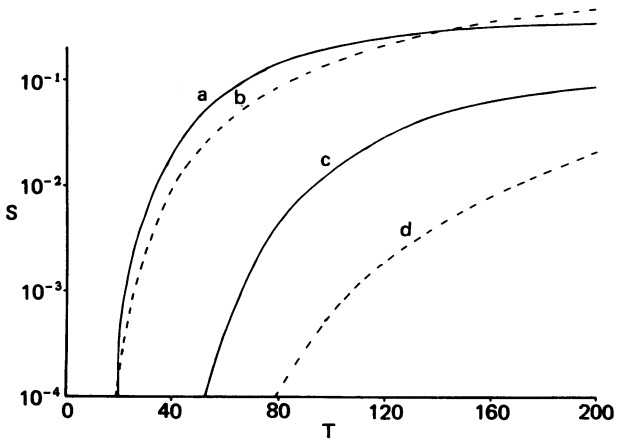

FIG. 1. Graph of the contribution, S, expressed as a proportion of $N\mu$, over T generations to the class of variants with over k copies, under either an initial population explosion or a steady growth pattern. (a) $m = 1.02, k = 100$; (b) $m = 1.2$ for 20 generations followed by $m = 1, k = 100$; (c) $m = 1.02, k = 400$; (d) $m = 1.2$ for 20 generations followed by $m = 1, k = 400$.

Table 2 the contributions computed from [4] for $k = 50$, 100, 400, and 1000 are tabulated for a variety of patterns of cyclic growth, i.e., constant rates of increase for a specified number of generations followed by a population crash. In addition to the three latter values, considered for reasons discussed in the previous paper (1), we have extended our analysis to the case $k = 50$, to include variants that bridge the gap between being present in a single extended family and being well-established private polymorphisms. We see from this table that, for example, under the cyclic pattern of increase at 20% per generation ($m = 1.2$), with a 32-fold decrease every 20 generations restoring the population to constant size (on average), the value of S is 0.135 for variants up to 200 generations old and having more than 400 replicates. Thus, we expect to see $0.135N\mu$ different variants in this class, and the standard deviation of the number of such variants is approximately $(0.135N\mu)^{1/2}$. (Since the distribution of S is Poisson, the confidence intervals are quite asymmetric.)

From Fig. 1 we see that for variants accumulating over the full 200 generations the contribution to the >400 class is greater with an initial population explosion than under a steady rate of increase, but that for the >100 class this is only so for variants under 140 generations old. Except in the initial expansion phase, which is not a disproportionate factor in the total contribution, variants are arising at a period of constant mean population size. However, due to the initial expansion, the population during the latter phase of its history is larger and so more such variants arise. For older variants this is the dominating factor.

The effects of different patterns of cyclic growth are substantial, but not an order-of-magnitude effect. Clearly, all contributions must increase with the number of generations considered. The proportion in the more highly replicated numbers increases with T, since older variants are, if not extinct, usually present in large numbers. Insofar as differences do exist between the different patterns of increase, the most important factor is the magnitude of the periodic disaster rather than the intervening rate of increase. Over large numbers of generations the contributions are in inverse ratio to the disaster magnitude. Over shorter time periods the major effect is of the value of m, and for $k = 1000$ this effect persists even over 400 generations, the populations with low m values rarely contributing to this class. That is, over shorter periods the "larger disaster" pattern will allow a few variants to become replicated in very large numbers. Over long periods, the "small disaster" pattern will allow more private polymorphisms.

APPLICATION OF THE THEORY

Our aim is to explain the observed data on the Amerindian private polymorphisms. From Table 1 we see we have six variants with over 50 copies, four variants with over 100 copies, three with over 400 copies, and two with more than 1000 copies. From the previous section we see that the values of the sum given by [4] over 400 generations for these four k values range from 0.385, 0.305, 0.151, and 0.008 to 1.90, 1.12, 0.276, and 0.111, with typical intermediate values being 0.8, 0.5, 0.2, and 0.07. Also from Table 1 we obtain a total N value over all systems and all populations of $2 \times 25 \times 37{,}296$ or 1,864,800. The expected number of private polymorphisms for a range of μ values are given in Table 3. The same results are given diagrammatically in Fig. 2 where the combination of $N\mu$ and S values giving the expectation in each class equal to the observed number are shown. For the particular case of Fig. 2, of variants arising over 400 generations having more than 100 replicates, we see that the data imply a value of $N\mu$ between 2.5 and 17.5, or a value of μ between 1.3×10^{-6} and 9.4×10^{-6}/locus per generation.

We see that a mutation rate of 10^{-6} requires us to find some explanation for the prevalence of the private polymorphisms other than the conditions that have been considered, while if the mutation rate were 5×10^{-5}, we would have, under these considerations, to explain why there are not many more. A

Table 2. Contribution, S, as a proportion of $N\mu$, to the class of variants having more than k copies for variants accumulating over T generations under various patterns of cyclic increase

m	Cycle length, generations	Disaster magnitude*	$T = 200$				$T = 400$			
			$k = 50$	$k = 100$	$k = 400$	$k = 1000$	$k = 50$	$k = 100$	$k = 400$	$k = 1000$
1.2	20	32	0.409	0.313	0.135	0.0480	0.519	0.420	0.224	0.111
1.15	30	58	0.313	0.234	0.119	0.0535	0.385	0.305	0.183	0.103
1.15	10	3.5	0.945	0.556	0.076	0.0040	1.393	0.948	0.256	0.045
1.1	40	41	0.380	0.274	0.126	0.0496	0.475	0.366	0.206	0.108
1.1	20	6.1	0.814	0.522	0.104	0.0107	1.146	0.824	0.276	0.069
1.08	40	20	0.516	0.367	0.135	0.0362	0.674	0.519	0.252	0.107
1.05	40	6.7	0.777	0.503	0.105	0.0115	1.097	0.796	0.276	0.072
1.02	100	7.1	0.783	0.471	0.098	0.0106	1.014	0.763	0.269	0.071
1.01	100	2.7	0.973	0.530	0.055	0.0017	1.504	0.982	0.230	0.031
1.00	—	—	1.115	0.511	0.020	0.0001	1.905	1.117	0.151	0.008

* The factor by which the population is decreased.

Table 3. Expected numbers of variants observed to have more than k copies, over 400 generations

	$S(k = 50$, 6 polymorphisms)			$S(k = 100$, 4 polymorphisms)			$S(k = 400$, 3 polymorphisms)			$S(k = 1000$, 2 polymorphisms)		
μ	0.385	0.8	1.90	0.305	0.5	1.12	0.151	0.2	0.276	0.0078	0.07	0.111
10^{-6}	0.72	1.49	3.54	0.57	0.93	2.09	0.28	0.37	0.51	0.01	0.13	0.21
3×10^{-6}	2.15	4.48	10.63	1.71	2.80	6.27	0.84	1.12	1.54	0.04	0.39	0.62
7×10^{-6}	5.03	10.44	24.80	3.98	6.53	14.62	1.97	2.61	3.60	0.10	0.91	1.45
10^{-5}	7.18	14.92	35.43	5.69	9.32	20.89	2.82	3.73	5.15	0.15	1.31	2.07
5×10^{-5}	35.90	74.59	177.16	28.44	46.62	104.43	14.08	18.65	25.73	0.73	6.53	10.35

The standard deviation of each expectation is approximately equal to the square root of the expectation.

mutation rate of the order of 7×10^{-6} gives the closest overall fit to the observed data, although the estimate of the number of variants with over 1000 copies is lower than the observed value, particularly in the case of an assumed constant m. In seeking the best fit with our data, we must bear in mind that in no case have we sampled the total tribe. Several of the low-frequency polymorphisms are quite localized in their distribution within the tribe in which they occur, and might have been missed had the tribe been sampled in a different way. Thus, our estimate of the number of private polymorphisms in these tribes is minimal as, then, would be the estimate of the mutation rate supporting them.

Without some periods of rapid expansion allowing a variant to become established, we will very, very rarely see variants with over 1000 copies. Both the tribes in which such variants are observed have probably doubled in size over the last century and are now relatively large in their size vis-à-vis the other surviving tribes of Tropical America. This has inflated the number of copies of the relevant variants, and perhaps only the Guaymi variant would have achieved 1000 copies under "normal" demographic conditions; an expectation of one such variant is given by patterns of intermediate cyclic variation. Implicit in the above treatment is the assumption of homogeneity within each variant class, an assumption which we believe to have a low probability of error in populations of this type.

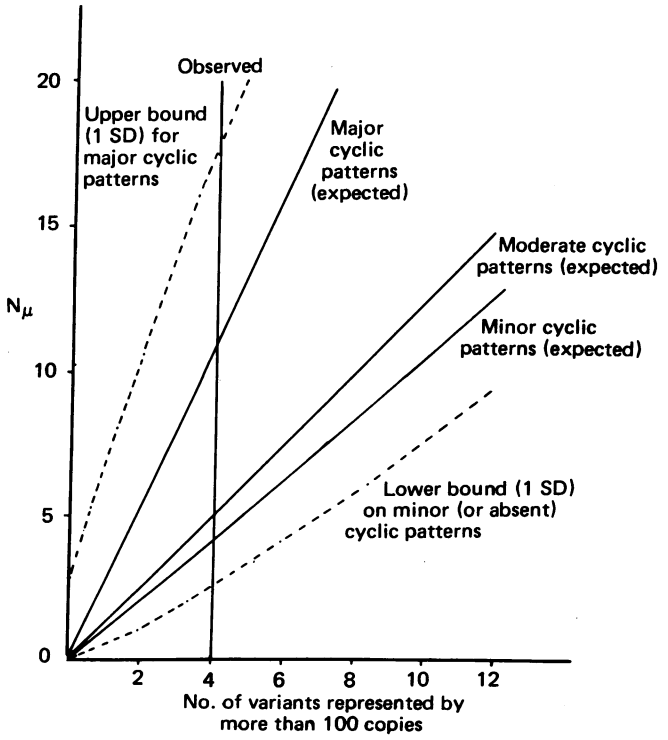

FIG. 2. Ranges of numbers of neutral variants arising over 400 generations in the class represented by more than 100 copies as a function of $N\mu$, under different demographic patterns. ($N\mu$ is the total number of new variants that will arise, and reach adulthood, in the current generation.) A minor cyclic pattern would, for example, be represented by a "disaster magnitude" of 2.7 and a major cyclic pattern, for example, by a "disaster magnitude" of 58.

DISCUSSION

In the foregoing calculations, we have tacitly assumed that the mutant allele increases or decreases with the waxing or waning of the population, i.e., that the m values for bearers of the mutant and of the normal allele are the same. This is an assumption of phenotypic neutrality for the heterozygous mutant allele. In a subsequent paper we will estimate the most probable average rate of mutation resulting in electrophoretic variants by a variety of approaches that use all the data on electrophoretic variants in these tribes. A significant discrepancy between that estimate and the estimate of 7×10^{-6}, which was found to be most consistent with the present data, would yield evidence for the operation of selection on traits of this type in these populations. As of now, the salient conclusion to be drawn from the present exercise is that given the breeding structure that we have established for Amerindian tribal populations (4, 5), the frequency of private polymorphisms in such populations can be explained by mutation rates of the order commonly considered for man coupled with genetic drift, without the obvious necessity of appealing to the action of selection.

However, the mutation rate just quoted, of 7×10^{-6}, applies only to a limited portion of the mutational spectrum at the average structural locus. In the case of the hemoglobin polypeptides, the great majority of electrophoretically detected variants can be shown to be due to single amino acid substitutions, which alter the net charge of the molecule. It is readily calculated that for a typical protein, only about one-third of all the possible one-step mutations that substitute one amino acid for another will result in a charge change. Furthermore, because of degeneracy in the genetic code, and depending on the specific protein, about 23–24% of all nucleotide substitutions will result in synonymous changes (5). Finally, recent studies on *Drosophila* suggest that mutations resulting in complete loss of enzyme activity are some 5–6 times more common than those resulting in electrophoretic variants in which enzyme activity is retained (6). While some of these "null" mutants may be due to single amino acid substitutions, it seems probable that others reflect more drastic changes in the molecule under study. Since the alleles responsible for "silent" amino acid substitutions and loss of enzyme activity must also be subject to genetic drift, one is led to postulate the existence of many more (electrophoretically silent) polymorphisms in these data than have thus far been detected, maintained *in toto* by a mutation rate considerably higher than that given above.

This work was supported in part by Energy Research and Development Contract EY-77-C-02-2828 and National Science Foundation Grant BMS-74-11823, and performed while E.A.T. was a Visiting Postdoctoral Scholar, Department of Human Genetics, University of Michigan Medical School, Ann Arbor, MI.

1. Thompson, E. A. & Neel, J. V. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 1442–1445.
2. Neel, J. V. (1978) *Am. J. Hum. Genet.* in press.
3. Neel, J. V. & Weiss, K. M. (1975) *Am. J. Phys. Anthrop.* **42**, 25–52.
4. Neel, J. V. (1977) *Annu. Rev. Genet.*, in press.
5. Drake, J. W. (1970) *The Molecular Basis of Mutation* (Holden-Day, Inc., San Francisco, CA).
6. Mukai, T. & Cockerham, C. C. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 2514–2517.