Proc. Nat. Acad. Sci. USA
Vol. 72, No. 8, pp. 3194–3196, August 1975
Genetics

Statistical analyses of *Drosophila* and human protein polymorphisms*

(population genetics/mutation pressure/theory of neutral protein variation)

TOMOKO OHTA

National Institute of Genetics, Mishima, 411 Japan

Communicated by Motoo Kimura, June 3, 1975

ABSTRACT By using the distribution function of allelic frequencies which was recently derived by Kimura and Ohta for the model of stepwise production of neutral alleles, the observed protein polymorphisms of *Drosophila* and man are tested for fit to the theory of neutral protein variation. The observed and theoretical distributions of alleles agree quite well except for the excess of rare alleles in the actual distributions. In human polymorphisms, the alleles with frequencies less than 1% are more numerous than expected, whereas in *Drosophila*, those with frequencies less than 10% are more numerous. It is pointed out that these results support my thesis that mutational pressure rather than balancing selection is the main cause for the maintenance of protein polymorphisms.

The biggest controversy in population genetics in recent years is the one between the neutral versus selection hypotheses of protein variations (1–5). Several attempts have been made to test whether the observed pattern of protein polymorphisms fits to the neutral theory (6–9). In these studies, the theoretical predictions are based on the model of neutral isoalleles originally proposed by Kimura and Crow (10). The model assumes that whenever mutation occurs it represents a new, not preexisting allele. On the other hand, most observations of protein polymorphisms rely on the electrophoretic methods that can only detect charge differences of a protein molecule. Because of such a limited detectability, it is now generally considered that the model assuming stepwise production of alleles is more appropriate to analyze observations on the electrophoretically detectable alleles. In this model, we assume that the entire sequence of allelic states can be expressed by integers, and that if an allele changes states by mutation, it moves either one step in the positive direction or one step in the negative direction in the allele space (4, 11).

Recently, we have been able to obtain a formula for the equilibrium distribution of the allelic frequencies under the assumption of stepwise production of neutral alleles (12). The distribution is given by

$$\Phi(x) = \frac{\Gamma(\alpha + \beta + 1)}{\Gamma(\alpha)\Gamma(\beta + 1)}(1 - x)^{\alpha-1}x^{\beta-1}, \quad [1]$$

where

$$\alpha = 4N_ev, \ \beta = \frac{1 + 4N_ev - \sqrt{1 + 8N_ev}}{\sqrt{1 + 8N_ev} - 1},$$

and N_e and v, respectively, denote the effective population size and the mutation rate (see ref. 12). This distribution has the meaning that $\Phi(x)dx$ represents the expected number of alleles whose frequencies in the population are in the range $(x \sim x + dx)$.

By using this distribution, it is now possible to make a more appropriate test of neutrality than was previously possible. In the present paper, I shall attempt statistical analyses to test if the observed allele distributions fit the above distribution. Recently the step model has been criticized by Johnson (13), who claims that it is not completely realistic and that the true situation must lie somewhere between the step model and the Kimura-Crow model. Therefore, I shall also include analyses using the distribution obtained under the Kimura-Crow model.

STATISTICAL ANALYSES

For the present analyses, I have chosen the data on the *Drosophila willistoni* group of Ayala *et al.* (14) and data on English human population of Harris *et al.* (15) because of their large sample sizes. Since the pattern of allele distribution is relatively uniform over the wide geographical range in *Drosophila* species (14, 16), global allele frequencies were used for the analysis of *Drosophila* data. The data of *Drosophila willistoni*, *D. tropicalis*, and *D. equinoxialis* were treated separately, whereas those of *D. paulistorum* and *D. nebulosa* were not included in the analyses because of small sample sizes. In the present analyses, the number of different alleles for a specified genê frequency range (say, between x and $x + \Delta x$) were counted in each species and summed over all loci. The corresponding theoretical number• per locus was computed by using formula [1] as follows;

$$\bar{n}_a(x, x + \Delta x) = \int_x^{x+\Delta x} \Phi(x)dx. \quad [2]$$

The distribution contains an unknown parameter N_ev (the product of the effective population size and the mutation rate) that has to be estimated from observations. According to Ewens (6), the number of alleles (n_a) actually found in the sample is a sufficient statistic for estimating N_ev. In actual populations, however, the value of n_a may be much influenced by the presence of very slightly deleterious mutations. Our preliminary study suggests their presence in many species (17). Therefore, we estimated N_ev from the average heterozygosity ($\bar{H}_e$) over all available loci for each species, since $\bar{H}_e$ is less dependent on rare mutant alleles. For the step model, the relation between $\bar{H}_e$ and N_ev is given by

$$\bar{H}_e = 1 - \frac{1}{\sqrt{1 + 8N_ev}} \quad [3]$$

* Contribution no. 1064 from the National Institute of Genetics, Mishima, Shizuoka-ken 411 Japan.

(11). Since the value of $\Phi(x)$ rises sharply as x approaches 0, the expected number of alleles with low frequencies may be much influenced by the sample size. For a particular locus of one species, the expected number of alleles with frequency less than 1% may be computed by the following formula;

$$\bar{n}_a(x < 0.01) = \int_{\frac{1}{2N}}^{0.01} [1 - (1 - x)^{2n}]\Phi(x)dx \quad [4]$$

(6, 12, 18), where n is the number of independent individuals in the sample and N is the actual size of the population. The value of n varies considerably from locus to locus in all species studied and this complicates the computation. I evaluated the expected number of alleles in this class for each enzyme locus of various sample sizes using formula [4] by assuming $2N = 10^6$ and summed over all loci in each species. Actually only little difference was found whether we assume $2N = 10^4$, 10^5, or 10^6 as long as $N \gg n$.

Table 1 shows the observed and the expected numbers of alleles for the four species studied. The table also includes the expected numbers of alleles computed using the infinite allele model of Kimura and Crow (10). It can be seen from the table that the theoretical predictions based on the neutral theory generally fit quite well to actual observations except for the low gene frequency classes (top 2 lines of Table 1). In fact, χ^2 values turned out to be very large for the frequency class of less than 1% in all four species. For the gene frequency class between 1% and 10%, the observed and the expected numbers of alleles agree well in man but not in *Drosophila*, χ^2 being significant for the latter species. It should also be noted that the deviations from the expectation based on the Kimura-Crow model are also significant.

DISCUSSION

The excellent agreement between observed and theoretical distributions of alleles, save for the excess of rare alleles, as shown above clearly indicates that the "balancing selection" is not the general mechanism for maintaining these alleles, particularly the low frequency ones. It rather indicates that the mutation pressure is playing the major role for their maintenance. Regarding more common alleles, the observed frequency distribution agrees remarkably well with the theoretical distribution both in man and *Drosophila*. Of course, this is not the unequivocal proof of the validity of the neutral hypothesis, since some sort of balancing selection could be invoked to explain the distribution equally well. However, the sharp rise of observed frequency distribution toward the origin could not be explained by any model of balancing selection.

The present analysis confirms our preliminary conclusion that too many alleles are actually segregating in the population as compared with the level of observed heterozygosity (17). Latter (19) also found the excess of rare alleles in several *Drosophila* species by using a different method. In fact, we can now see the nature of this deviation; rare alleles whose frequency is less than 1% are roughly five times as numerous as expected from the balance between mutational input and random extinction of strictly neutral alleles. This tendency seems to exist both in *Drosophila* and in man. The difference between the human and *Drosophila* polymorphisms is evident in the next frequency class; the alleles whose frequencies lie in the range 0.01 ~ 0.1 are three to four times as numerous as expected from the neutral theory in the case of *Drosophila*, whereas the number of alleles in this class is just as expected in human population.

There are two possibilities which might result in the excess of rare alleles by recurrent mutations: (*1*) populations studied have rapidly expanded recently, and (*2*) rare alleles are on the average very slightly deleterious and they are maintained essentially by mutation-selection balance. The first possibility, as pointed out by M. Nei (personal communication), predicts that these rare alleles will increase their frequencies in future generations. Let us examine this case using a simple model. Suppose that the population expanded very rapidly under the condition $N_ev \gg 1$. We assume that before expansion there was a single allele (A_0). After expansion, the number of alleles increases by mutation and we again assume the following step model:

$$\begin{array}{ccccccccc} & v/2 & & v/2 & & v/2 & & v/2 & \\ \ldots A_{-2} & \rightleftarrows & A_{-1} & \rightleftarrows & A_0 & \rightleftarrows & A_1 & \rightleftarrows & A_2 \ldots \\ x_{-2} & & x_{-1} & & x_0 & & x_1 & & x_2 \end{array}$$

Table 1. The observed and expected numbers of alleles of various frequency classes

	D. willistoni			*D. tropicalis*			*D. equinoxialis*			Human (English)		
		Expected			Expected			Expected			Expected	
Allele frequency	Ob-served	Step	K-C	Ob-served	Step	K-C	Ob-served	Step	K-C	Ob-served	Step	K-C
Estimated value of N_ev. . .		0.060	0.054		0.045	0.042		0.054	0.049		0.020	0.020
< 0.01	85*	11.15	18.79	41*	8.65	12.92	56*	10.29	16.34	59*	10.84	14.07
0.01 – 0.1	39*	12.56	15.92	29*	9.91	12.01	44*	11.75	14.46	8	7.08	8.02
0.1 – 0.2	8	4.84	5.25	5	3.71	3.98	3	4.48	4.78	0	2.51	2.69
0.2 – 0.3	2	3.31	3.40	1	2.52	2.59	2	3.06	3.10	3	1.69	1.77
0.3 – 0.4	1	2.72	2.70	1	2.07	2.07	1	2.51	2.47	1	1.39	1.43
0.4 – 0.5	2	2.47	2.39	2	1.89	1.85	4	2.28	2.19	2	1.27	1.30
0.5 – 0.6	2	2.40	2.29	2	1.85	1.79	1	2.23	2.11	3	1.26	1.28
0.6 – 0.7	2	2.51	2.36	1	1.95	1.87	2	2.36	2.19	1	1.35	1.37
0.7 – 0.8	0	2.86	2.67	2	2.26	2.15	1	2.68	2.49	2	1.61	1.62
0.8 – 0.9	2	3.84	3.57	3	3.12	2.95	4	3.63	3.37	0	2.34	2.34
0.9 – 1.0	23	19.19	19.54	21	20.86	21.04	21	20.02	20.43	37	36.79	36.64

The expected values were obtained by integrating the distribution function numerically. The observed values are the sum of all loci in each species. For details, see *text*. K-C refers to the Kimura-Crow model.

* Significantly larger than expected.

Then it can be shown that the allele frequencies as functions of time are given by the following formulae,

$$x_0(T) = \sum_{k=0}^{\infty} \frac{\binom{2k}{k}(-1)^k}{k!} T^k$$

$$= 1 - 2T + 3T^2 - \frac{10}{3}T^3 + \ldots$$

$$x_{\pm 1}(T) = \sum_{k=1}^{\infty} \frac{\binom{2k}{k+1}(-1)^{k-1}}{k!} T^k$$

$$= T - 2T^2 + \frac{5}{2}T^3 - \ldots$$

$$x_{\pm 2}(T) = \sum_{k=2}^{\infty} \frac{\binom{2k}{k+2}(-1)^{k-2}}{k!} T^k$$

$$= \frac{1}{2}T^2 - T^3 + \ldots, \text{etc.},$$

where $T = vt/2$ in which t represents the number of generations after the population expansion started and $x_i(T)$ stands for the allele frequencies at time T. If T is small (for example, 10^{-2}), we expect a polymorphic pattern where a common allele is surrounded by rare alleles. Then, the observed difference between human and *Drosophila* may be explained by assuming that T is much smaller in humans than in *Drosophila*.

On the other hand, the alternative possibility of slight negative selection predicts that *Drosophila* populations may be near to the true equilibrium between mutation and selection (20, 21), whereas in human populations such an equilibrium has not yet been reached. Observed relative uniformity of heterozygosity over many species and the heterogeneity of electrophoretic alleles with respect to heat-sensitivity (22) seem to suggest the latter possibility (23). We cannot discriminate, at the moment, either one of the two possibilities from the observed frequency distributions of the alleles.

The author would like to express her deep gratitude to Dr. Motoo Kimura for his kind advice and encouragement throughout the course of this work. Thanks are also due to him and Dr. Masatoshi Nei for many stimulating discussions.

1. Kimura, M. (1968) *Nature* **217,** 624–626.
2. King, J. L. & Jukes, T. H. (1969) *Science* **164,** 788–798.
3. Kimura, M. & Ohta, T. (1971) *Nature* **229,** 467–469.
4. Kimura, M. & Ohta, T. (1973) *Genetics* (Sup.) **73,** 19–35.
5. Lewontin, R. C. (1974) *The Genetic Basis of Evolutionary Change* (Columbia University Press, New York and London).
6. Ewens, W. J. (1972) *Theor. Pop. Biol.* **3,** 87–112.
7. Johnson, G. B. (1972) *Nature New Biol.* **237,** 170–171.
8. Yamazaki, T. & Maruyama, T. (1972) *Science* **178,** 56–58.
9. Nei, M. & Roychoudhury, A. K. (1974) *Genetics* **76,** 379–390.
10. Kimura, M. & Crow, J. F. (1964) *Genetics* **49,** 725–738.
11. Ohta, T. & Kimura, M. (1973) *Genet. Res.* **22,** 201–204.
12. Kimura, M. & Ohta, T. (1975) *Proc. Nat. Acad. Sci. USA,* **72,** 2761–2764.
13. Johnson, G. B. (1974) *Genetics* **78,** 771–776.
14. Ayala, F. J., Tracey, M. L., Barr, L. G., McDonald, J. F. & Perez-Salas, S. (1974) *Genetics* **77,** 343–384.
15. Harris, H., Hopkinson, D. A. & Robson, E. B. (1974) *Ann. Hum. Genet.* **37,** 237–253.
16. Prakash, S., Lewontin, R. C. & Hubby, J. L. (1969) *Genetics* **61,** 841–858.
17. Ohta, T. & Kimura, M. (1974) *Genetics* **76,** 615–624.
18. Kimura, M. (1974) in *Lectures on Mathematics in the Life Sciences* (American Mathematical Society), Vol. 7, pp. 1–23.
19. Latter, B. D. H. (1975) *Genetics* **79,** 325–331.
20. Ohta, T. & Kimura, M. (1975) *Amer. Nat.,* **109,** 137–145.
21. King, J. L. & Ohta, T. (1975) *Genetics,* **79,** 681–691.
22. Bernstein, S. C., Throckmorton, L. H. & Hubby, J. L. (1973) *Proc. Nat. Acad. Sci. USA* **70,** 3928–3931.
23. Ohta, T. (1974) *Nature* **252,** 351–354.