Am J Hum Genet 30:262–270, 1978

Ethnic Variation of Genetic Disease: Roles of Drift for Recessive Lethal Genes

DIANE WAGENER,[1,2,3] LUIGI L. CAVALLI-SFORZA,[1] AND RICHARD BARAKAT[2]

INTRODUCTION

The importance of the rare recessive gene to medical genetics is well known [1]. The possible roles of nonequilibrium, hitchhiking, and epistatic models to explain variations of diseases between ethnic groups have also been studied [2]. In this paper, we show that deviations between groups observed for phenylketonuria (PKU), cystic fibrosis, and Tay-Sachs are possibly the result of genetic drift in restricted populations. We use the estimates of mean gene frequencies for PKU ($6 \cdot 10^{-3}$) and cystic fibrosis ($1.02 \cdot 10^{-2}$) as given in tables 1 and 2 in reference [2].

Considerable attention has recently been focused on the ethnic variation of Tay-Sachs disease, with some investigators suggesting either founder effect [4–7] or selection in the form of heterozygote advantage [8–13] to explain the high frequency in certain populations. The data from several surveys listed in table 1 demonstrate the increased frequency of the Tay-Sachs gene among Ashkenazi Jews; those from Israel, although the smallest sample, indicate the variation between Ashkenazi groups.

Myrianthopoulos et al. [8–11], looking for a heterozygote advantage in Tay-Sachs among the Ashkenazim, found a moderate increase in fitness in the sibs of the affected and suggested tuberculosis (TB) as a possible selective factor imparting advantage to heterozygotes. Naturally, this hypothesis requires further explanation of the failure of other non-Jewish groups, living under very similar conditions, to show an increase in Tay-Sachs frequency. The heterozygotic advantage for Tay-Sachs predicted from the assumption of equilibrium is 3%–6% [11–12]. Even if there is an indication of selective advantage of heterozygotes from demographic data of relatives of probands, the observed differences are far from significant.

Rao and Morton [18] suggested that drift could explain the ethnic variation for Tay-Sachs and cystic fibrosis. However, they assumed a beta distribution of gene frequencies among populations which is strictly valid only in the absence of selection

Received May 28, 1976; revised November 22, 1977.

This work was supported in part by grant 1-F32-GM-05455-01 (at Harvard) from the National Institutes of Health and grant AT(04)-326-PA-433 (at Stanford) from the Atomic Energy Commission.

[1] Department of Genetics, Stanford University, Stanford, California 94305.

[2] Division of Engineering and Applied Physics, Harvard University, Cambridge, Massachusetts 02138. R. Barakat is also associated with Bolt Beranek and Newman, Inc., Cambridge, Massachusetts 02138.

[3] Present address: Western Psychiatric Institute and Clinic, University of Pittsburgh School of Medicine, 3811 O'Hara Street, Pittsburgh, Pennsylvania 15162.

TABLE 1

TAY-SACHS GENE FREQUENCIES FOR SEVERAL POPULATIONS

Region and Source	No. Births	Incidence	Gene Frequency
Israel, Ashkenazi [14]	85,000	1/5,000	.0141
New York City, Ashkenazi [15]	450,000	1/8,333	.0110
United States, Ashkenazi [7]	352,000	1/6,250	.0127
New York City, Ashkenazi [16]	412,500	1/4,350	.0152
Sephardic Jews [17]	...	1/588,250	.0013
New York City, non-Jews [15]	1,350,000	1/454,550	.0015
United States, non-Jews [7]	15,461,000	1/588,250	.0013
New York City, non-Jews [16]	1,012,500	1/384,600	.0016
Israel, non-Ashkenazi [14]	77,000	1/39,500	.0050

(or with weak selection, if any, in favor of the heterozygote). In the present case, there is strong selection against homozygote recessives. Wright [19, 20] has given the appropriate theoretical equilibrium distribution which we will henceforth call the lethal distribution.

EQUILIBRIUM UNDER DRIFT

Following the analysis of Wright, the distribution of the gene frequency, x, of a recessive lethal gene under mutation and selection in a finite population at equilibrium for mean and variance is given by

$$\phi(x) = Cx^{4Nu-1}(1-x^2)^{2N}(1-x)^{-1}, \tag{1}$$

where u is the mutation rate of the recessive allele, and N is the effective population size for the diploid population. The constant C is a normalizing factor such that the total probability is unity. Recalling that most of the probability mass is near $x = 0$, equation (1) can be rewritten, introducing the parameter $\theta = 4Nu$, as

$$\phi(x) \approx Cx^{\theta-1}e^{-2Nx^2} \tag{2}$$

$$\approx \frac{2(2N)^{\frac{1}{2}\theta}}{\Gamma(\frac{1}{2}\theta)}x^{\theta-\frac{1}{2}}e^{-2Nx^2}. \tag{3}$$

Here $\Gamma(\cdot)$ is the standard gamma function, and we have made the approximation

$$C = \frac{2(2N)^{\frac{1}{2}\theta}}{\Gamma(\frac{1}{2}\theta)} \tag{4}$$

by assuming that the upper bound for x is ∞ rather than 1. There is negligible error here because the probability mass of $\phi(x)$ is concentrated near $x = 0$.

Given the probability density function $\phi(x)$, we calculate the probability $P(x \geq x_T)$ that a population chosen at random will have a gene frequency x greater than or equal to a specified value x_T. Here the probability is approximated by

$$P(x \geq x_T) = C\int_{x_T}^{1} x^{\theta-1}e^{-2Nx^2}dx = 1 - C\int_{0}^{x_T} x^{\theta-1}e^{-2Nx^2}dx. \tag{5}$$

The integral in equation (5) can be found by using the incomplete gamma function

$$\gamma(x,p) = \int_0^x e^{-t} t^{p-1} dt \ ,$$

and this leads to

$$P(x \geq x_T) = 1 - \gamma(2Nx_T^2, \tfrac{1}{2}\theta)/\Gamma(\tfrac{1}{2}\theta) \ . \tag{6}$$

Computer routines are available for the evaluation of the incomplete gamma. We expressed this function in terms of the confluent hypergeometric function (equation 6.5.12 in reference [21]) when x_T is small. When x_T is large, its complement may be written as an asymptotic series in $x = 2Nx_T^2$ (equation 6.5.32 [21]).

The probability that a deviant gene frequency x is greater than a given value x_T can thus be calculated. It is somewhat more illuminating to plot $P(x \geq x_T)$ as a function of $T = x_T/\bar{x}$ (where $\bar{x}$ is the expected gene frequency, discussed below) than of x_T itself. In figures 1 and 2, the probabilities for various values of θ are graphed. From figure 2, it is apparent that large deviations are probable if the number of mutations per generation ($2Nu$) is small (i.e., $\theta < \frac{1}{2}$). Note that these probabilities are functions of θ only and do not depend directly on the effective population size N.

The moments of degree r for the gamma distribution of equation (5) are

$$E(x^r) = \Gamma(\tfrac{1}{2}\theta + \tfrac{1}{2}r)/[(2N)^{\frac{1}{2}r}\Gamma(\tfrac{1}{2}\theta)] \ , \tag{7}$$

so that the expected gene frequency $\bar{x}$ is $\Gamma(\frac{1}{2}\theta + \frac{1}{2})/[(2N)^{\frac{1}{2}}\Gamma(\frac{1}{2}\theta)]$ which depends on both θ and N. Some representative values are listed in table 2.

Therefore, given the value of $\bar{x}$ and the order of magnitude of the estimated N for a population, the corresponding values of θ can be obtained from table 2, and the probabilities of certain deviants can be determined from figures 1 and 2. For instance, consider cystic fibrosis, a disease of European Caucasians, for which the Caucasian gene frequency ($x = .01616$) deviates as much as five times from the Oriental gene frequency ($x = .0033$). The greater Caucasian gene frequency is very high for lethal gene frequencies maintained by mutation and selection. The higher frequency, therefore, may be a deviation from a baseline gene frequency, comparable to the Oriental gene frequency. Consequently, we estimate the probability of a deviation five times a mean gene frequency of .0033. As an estimate of N, we choose effective population sizes of 10^4, that is, census sizes of 30,000 or higher. Although the population of the whole of England before the transition to agriculture was estimated to be less than 10,000 people [22], subsequent population changes have diluted the effects of the initial population size. In table 3, the probabilities of deviations of five times the mean gene frequency are given with several mean gene frequencies and several θ values. Also, the corresponding probabilities have been interpolated from table 2 in Rao and Morton for the beta distribution. From table 2, given $N = 10{,}000$ and mean gene frequency of .0033, the corresponding value of θ is .8. The probabilities derived

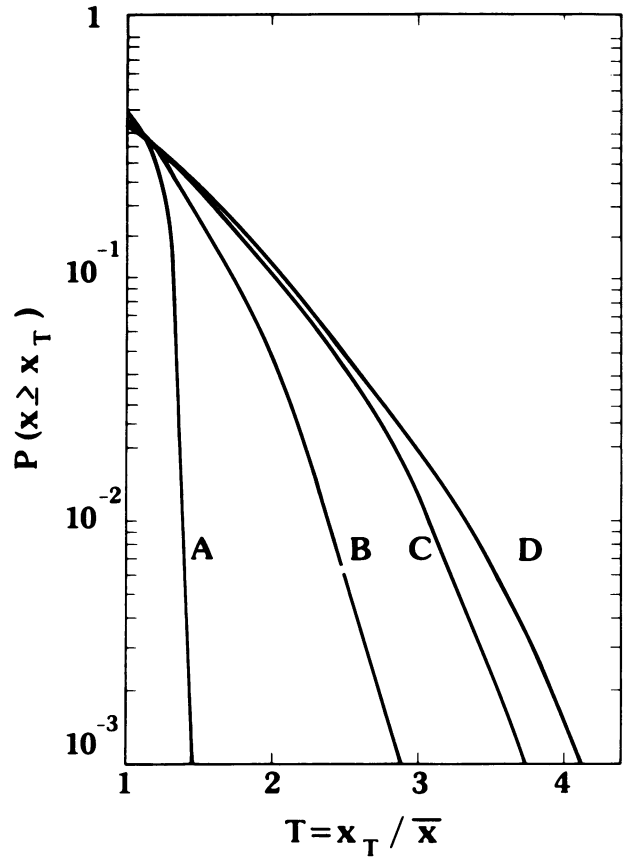

FIG 1.—$P(x \geq x_T)$ as a function of T for various values of θ greater than or equal to unity: A, $\theta = 11$; B, $\theta = 2$; C, $\theta = 1.1$; D, $\theta = 1.01$. Note that x_T is expressed as a multiple (T) of x.

from Wright's lethal distribution are not only somewhat smaller than those obtained from the beta distribution, but these values also fall much more rapidly as the expected gene frequency increases. For instance, in table 3 we note that as $\bar{x}$ increases by 50% (from .002 to .0028), the decrease in probability is a factor of six for the lethal distribution, but only three for the beta distribution. Note from table 3 that we are, in fact, concerned with three parameters. Given the estimated mean gene frequency and population size, a choice of θ implies a choice of mutation rate.

Using the procedure above, a probability of, say, .02 indicates that such a large deviance is improbable under conditions generating that distribution for a random observation from that distribution. In accordance with standard convention, a probability of less than 5% and certainly for less than 1% indicates that there may be a significant departure from the assumptions generating the distribution. But we have to consider that the diseases we are examining are not random observations. In fact, they are extreme outliers. These diseases may represent the extreme deviants of all genetic diseases known today, which number in the thousands [1]. A chance deviation of the

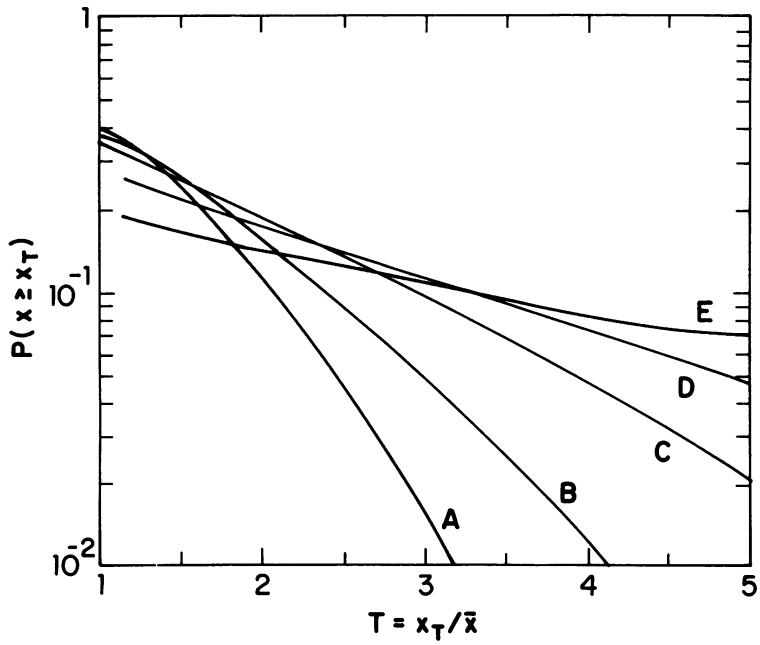

FIG. 2.—$P(x \geq x_T)$ as a function of T for various values of θ less than or equal to unity: A, $\theta = 1$; B, $\theta = .6$; C, $\theta = .4$; D, $\theta = .2$; and E, $\theta = .1$.

order one in a 1,000 for one of these diseases would be expected because of their having been selected, if they indeed are the extreme outliers.

In the case of cystic fibrosis, with $\bar{x} = .0033$ the corresponding θ value for $N = 10{,}000$ is .8. A deviation of $x > 5\bar{x}$, as observed, has a probability equal to .001 (table 3). This is certainly in agreement with the fact that this disease is an outlier out of perhaps thousands of diseases (see also Wright and Morton [23]). The agreement with the probability figure obtained by Wright and Morton ($P = .0012$) is probably coincidental since we used different data and an unrelated method, but the agreement is better than that with the figure obtained by Rao and Morton [18] using the beta distribution (approximately $P = .01$, extrapolating from their tables). It should be noted, however, that all of these methods imply a decision on N. If N were smaller (table 3, $N = 1{,}000$, and $\theta = .2$), the probability would be higher (.047).

Turning our attention to PKU, we want to find the probability of deviations of only two or three times a mean gene frequency of 6×10^{-3} in populations of size 10,000. In table 3, note that as the mean gene frequency increases, the probabilities of deviants rapidly decrease, so that for large population sizes and mean gene frequencies, deviations even as large as two times are not probable. From table 2, we find θ

TABLE 2

EXPECTED GENE FREQUENCIES FOR GIVEN EFFECTIVE POPULATION SIZES AND θ VALUES

θ*	$N = 10^3$	$N = 10^4$	$N = 10^5$
.1	.0019	.0006	.0002
.2	.0035	.0011	.0004
.6	.0088	.0028	.0009
.8	.0109	.0034	.0010
.9	.0118	.0037	.0012
1.0	.0127	.0040	.0013
1.1	.0136	.0046	.0014
2.0	.0199	.0063	.0020
11.0	.0514	.0162	.0051

NOTE.—Choices of N and θ imply a choice of mutation rate. The expected gene frequencies with mutation rates of comparable magnitude are enclosed in the table by solid lines. From upper right to lower left, the grouped mutation rates are of orders 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, and 10^{-2}.

* $\theta = 4Nu$, where u is the mutation rate.

approximately 2.0. For $\theta = 2.0$, $\bar{x}$ is .0063 (for $N = 10{,}000$ as before), the probability of deviants with twice the gene frequency is .04, a little higher than for cystic fibrosis. This case is not covered in Rao and Morton's tables. The high frequency of PKU in Caucasians could be due to drift even more easily than cystic fibrosis.

TABLE 3

ESTIMATES FOR PROBABILITY OF DEVIATIONS

θ	MUTATION RATE (u)	MEAN GENE FREQUENCY	PROBABILITIES		
			$x_T = 2\bar{x}$	$x_T = 5\bar{x}$	$x_T = 5\bar{x}$*
		A.) $N = 10{,}000$			
.2	5×10^{-6}	.0011	.16	.05	>.013
.4	1×10^{-5}	.0020	.18	.013	.033
.6	1.5×10^{-5}	.0028	.16	.002	.013
.8	2×10^{-5}	.0034	.13	.001	. . .
1.1	2.7×10^{-5}	.0046	.10	.000	. . .
2.0	5×10^{-5}	.0063	.04	.000	. . .
		B.) $N = 1{,}000$			
.1	2.5×10^{-5}	.0019	.143	.066	.053
.2	5×10^{-5}	.0035	.174	.047	. . .
.4	1×10^{-4}	.0063	.175	.013	. . .

NOTE.— . . . = parametric values fall outside the ranges given in the tables of Rao and Morton [18].

* Interpolated from Rao and Morton [18], derived from beta distribution.

For Tay-Sachs, we consider a disease for which only one restricted population has characteristic gene frequencies which deviate substantially from the other populations in general. Historical information on population sizes of Ashkenazim has been recently collected by Fraikor [24–25]. She indicates that three "population bottlenecks" have occurred: (1) in the fifth through eighth centuries; (2) in the thirteenth through fifteenth centuries (between the crusades); and (3) in the seventeenth and eighteenth centuries. During these times the population effective size, taken as one-third the total census size, may have been not far from 10,000 or even less. Historical accounts suggest that population consisted of small isolated bands for substantial periods of time. The effective population size, therefore, was somewhat less than 10,000.

From gene markers different from the disease markers studied here, a population size of comparable order of magnitude may also be estimated [26]. A population size of about 10,000, therefore, is estimated by an independent method. The present large population size is the consequence of recent increases.

It is therefore reasonable, in discussing the frequency of the Tay-Sachs gene in Ashkenazi populations, to choose effective population sizes of $N = 10^3$ and 10^4. Taking first the value $N = 10^3$, equation (7) [table 2] indicates that it is necessary to choose $\theta = .08$ to arrive at a mean gene frequency of .0015, the observed non-Ashkenazi mean value. With these values, equation (6) shows that the probability of a frequency greater than or equal to the observed Ashkenazi value (.0133) is about .031. The population size of $N = 5{,}000$ is consistent with historical evidence mentioned above, giving a probability of deviation of .007 ($\theta \simeq .183$). Similarly, for $N = 10{,}000$, the probability is .12% and for $N = 20{,}000$, the probability is .004%.

Note how fast the probabilities fall with increasing population size. However, the probability of deviation depends on the assumed mean gene frequency of .0015, which is in agreement with estimates of non-Jews and Sephardic Jews living outside of Israel. The data on non-Ashkenazi Jews living in Israel showed an increased gene frequency. In this case, the deviation of the Ashkenazi gene frequency from the non-Ashkenazi would be less and, consequently, the probability of deviation greater.

These probabilities estimated above are less than the figure of 5%, which is the standard threshold employed in significance tests. But for reasons already discussed, we see no special merit for this particular application of the standard threshold. Besides being arbitrary, this threshold ordinarily applies to random samples, which these diseases are not.

DISCUSSION

The choice between random genetic drift and selection is always a problem in the absence of good measurements of fitness, which are very difficult to obtain. For each of the three diseases considered here, we lack solid evidence that selection is at work, and we find that the deviations are really not so extreme if drift is responsible. Of the three diseases, Tay-Sachs is the one for which there is more demographic evidence and better agreement with the hypothesis of drift. PKU is intermediate, and cystic fibrosis is perhaps the one that requires the most extreme chance deviation if drift is responsible for ethnic variation.

It should, of course, be made clear that evidence in favor of drift is in general weak. Drift can be accepted (1) if there is absence—or insufficiency—of proofs in favor of natural selection in the individual case being considered; and (2) if there is a minimum of demographic data available which is in basic agreement with the event being due to drift.

One should stress the lack of available data. The insufficiency of evidence for selection for the heterozygotes may be only due to the absence of adequate investigation, which is made even more difficult for cystic fibrosis and PKU by the lack of easy and fully reliable tests for heterozygotes. The present conclusion in favor of drift will have to be reversed if specific facts in favor of natural selection are demonstrated in the future. The fact that drift is compatible with the observations is no proof that it is the cause of it.

SUMMARY

Using Wright's distribution of gene frequencies for a recessive lethal gene, a method is given to analyze the probability that any particular gene frequency is greater than a given threshold gene frequency. The method is introduced to analyze the plausibility of drift for explaining observed data.

REFERENCES

1. McKusick VA: *Mendelian Inheritance in Man: Catalogs of Autosomal Dominant, Autosomal Recessive, and X-Linked Phenotypes*, 3d ed. Baltimore, Johns Hopkins Univ. Press, 1971
2. Wagener DK, Cavalli-Sforza LL: Ethnic variation in genetic disease: possible roles of hitchhiking and epistasis. *Am J Hum Genet* 27:348–364, 1975
3. Chase GA, McKusick VA: Founder effect in Tay-Sachs disease. *Am J Hum Genet* 24:339–340, 1972
4. McKusick VA: Clinical genetics at a population level: the ethnicity of disease in the United States. *Ala J Med Sci* 3:408–424, 1966
5. McKusick VA: Ethnic distribution of disease in non-Jews. *Isr J Med Sci* 9:1375–1382, 1973
6. McKusick VA: The ethnic distribution of disease in the United States. *J Chronic Dis* 20:115–118, 1967
7. Myrianthopoulos NC: Some epidemiologic and genetic aspects of Tay-Sachs disease, in *Cerebral Sphingolipidosis: A Symposium on Tay-Sachs Disease and Allied Disorders*, edited by Aronson SM, Volk BW, New York, Academic Press, 1962, pp 359–374
8. Myrianthopoulos NC, Aronson SM: Population dynamics of Tay-Sachs diseases. I. Reproductive fitness on selection. *Am J Hum Genet* 18:313–327, 1966
9. Myrianthopoulos NC, Aronson SM: Population dynamics of Tay-Sachs disease. II. What confers the selective advantage upon the Jewish heterozygote?, in *Proceedings 4th International Symposium on Sphingolipidosis*, New York, Plenum Press, 1972, pp 561–570
10. Myrianthopoulos NC, Naylor AF, Aronson SM: Tay-Sachs disease is probably not increasing. *Nature* 227:609, 1970
11. Conneally PM, Merritt AD, Yu P: Cystic fibrosis: population genetics. *Tex Rep Biol Med* 31(4):639–650, 1973
12. Shaw RF, Smith AP: Is Tay-Sachs disease increasing? *Nature* 224:1214–1215, 1969
13. Fine M, Himmelfarb M, Jelendo M: *American Jewish Yearbook 1972*. New York, Jewish Publishing Society of America, 1973

14. GOLDSCHMIDT E, LENX R, MERLIN S, RONEN A, RONEN I: The frequency of the Tay-Sachs gene in the Jewish communities of Israel, presented at 25th Annual Meeting of the Genetics Society of America, Storr, Conn., 1956
15. KOZINN PJ, WINER H, COHEN P: Infantile amaurotic idiocy. *J Pediatr* 51:58–64, 1957
16. ARONSON SM: Epidemiology, in *Tay-Sachs Disease*, edited by VOLK BW, New York, Grune and Stratton, 1964, pp 118–153
17. SHAW RF, SMITH AP: Is Tay-Sachs disease increasing? *Nature* 224:1214–1215, 1969
18. RAO DC, MORTON NE: Large deviations in the distribution of rare genes. *Am J Hum Genet* 25:594–597, 1973
19. WRIGHT S: The distribution of gene frequencies in populations. *Genetics* 23:307–320, 1937
20. WRIGHT S: *Evolution and Genetics of Populations*, vol. 2, *The Theory of Gene Frequencies*. Chicago, Univ. Chicago Press, 1969
21. ABRAMOWITZ M, STEGUN IA: *Handbook of Mathematical Functions*. New York, Dover, 1965
22. CLARK JGD: *Star Carr: A Case Study in Bioarcheology*, module no. 10. Menlo Park, Calif., Addison-Wesley, 1972
23. WRIGHT S, MORTON NE: Genetic studies on cystic fibrosis in Hawaii. *Am J Hum Genet* 20:157–169, 1968
24. FRAIKOR AL: An anthropological analysis of Tay-Sachs disease: genetic drift among Ashkenazim Jews. Ph.D. thesis, Denver, Univ. Colorado, 1973
25. FRAIKOR AL: Tay-Sachs disease: genetic drift among Ashkenazim Jews. *Soc Biol* 24:117–134, 1977
26. CARMELLI D, CAVALLI-SFORZA LL: The genetic origin of the Jews: a multivariate approach. In preparation

New Editor

Beginning July 1, 1978, all manuscripts and correspondence concerning editorial matters should be directed to the new editor, Dr. David E. Comings, Department of Medical Genetics, City of Hope Medical Center, 1500 E. Duarte Road, Duarte, California 91010.