Am J Hum Genet 25:594–597, 1973

Large Deviations in the Distribution of Rare Genes

D. C. RAO[1] AND N. E. MORTON[1]

A gene rare in the species is sometimes found to have an appreciable frequency in a restricted population, suggesting either genetic drift, founder effect, or local selective advantage. Recently, considerable attention has been drawn to this logical disjunction through controversy on Tay-Sachs disease [1, 2], which is due to a recessive gene relatively common in Ashkenazic Jews. While heterozygote advantage is a possible explanation, the evidence for it is not compelling [3–5]. Here we show that such large deviations are not improbable, and they could very well occur due to chance alone.

METHOD

Let us consider the general population mean gene frequency for Tay-Sachs disease, $\bar{Q} = .0013$, and the high observed gene frequency of Ashkenazic Jews, $q = .0126$ [2]. Let Q denote, in general, the gene frequency. We shall now treat Q as a random variable taking values between 0 and 1, with mean $\bar{Q}$ and a certain variance σ^2, which we take as $\sigma^2 = F\bar{Q}\ (1 - \bar{Q})$, where F is an inbreeding coefficient describing population subdivision. Our problem is to obtain the probability of the event $(Q \geqslant q = .0126)$ for given values of $\bar{Q}$ and σ^2 (i.e., for given values of F, with $\bar{Q} = .0013$). Regardless of the distribution of Q, we can obtain an upper bound for this probability. This is given by a generalization of Tchebyshev's inequality, called Cantelli's inequality (for example, see [6]): $P\ (Q \geqslant q) \leqslant \sigma^2/(\sigma^2 + t^2) = C$, say, $(t = q - \bar{Q})$. Values of C for different F are given in table 1. It is clear that these values are rather large, but do not give us any idea as to how small the probabilities actually are.

To get an idea about the exact probabilities, we now assume Q to follow a β distribution given by $f(x;a,b) = [\Gamma(a+b)/\Gamma(a)\Gamma(b)]x^{a-1}\ (1-x)^{b-1}$, $0 \leqslant x \leqslant 1$, where the two parameters of the distribution are $a = \bar{Q}(1-F)/F$ and $b = (1-\bar{Q})(1-F)/F$. Letting $I_q(a,b) = \int_0^q f(x;a,b)\ dx$, the probability of interest is given by $P(Q \geqslant q) = 1 - I_q(a,b) = p$, say. Available tables on the β distribution do not give values of $I_q(a,b)$ for the range of a and b we encounter here. Gen-

Received February 27, 1973; revised June 6, 1973.

PGL paper no. 103. This work was supported by grant 1 R01 HD06003 from the U.S. National Institutes of Health.

[1] Population Genetics Laboratory, University of Hawaii, Honolulu, Hawaii 96822.

TABLE 1

VALUES OF C, p, $P_N(1)$ AND $P_N(5)$ FOR SEVERAL VALUES OF INBREEDING FOR TAY-SACHS DISEASE AND CYSTIC FIBROSIS LOCI

Locus	F	C	p	$P_N(1)$		$P_N(5)$	
				$N = 500$	$N = 2{,}000$	$N = 500$	$N = 2{,}000$
Tay-Sachs disease	0.002	0.0199	0.0006	0.2771	0.7269	0.0000	0.0106
	0.005	0.0484	0.0095	0.9912	1.0000	0.5119	1.0000
	0.01	0.0923	0.0218	1.0000	1.0000	0.9836	1.0000
	0.05	0.3370	0.0270	1.0000	1.0000	0.9974	1.0000
Cystic fibrosis	0.0009	0.0177	0.0000	0.0059	0.0233	0.0000	0.0000
	0.005	0.0908	0.0179	0.9999	1.0000	0.9429	1.0000

NOTE.—See text for explanation.

erally, a is small and b is large for reasonable magnitudes of inbreeding. We compute $I_q(a,b)$ as follows [7, eq. 26.5.4]:

$$I_q(a,b) = \frac{q^a(1-q)^b}{aB(a,b)}\left[1 + \sum_{n=0}^{\infty} \frac{B(a+1,n+1)}{B(a+b,n+1)} q^{n+1}\right]$$

$$= \frac{q^a(1-q)^b\Gamma(a+b)}{\Gamma(a+1)\Gamma b}\left[1 + \sum_{n=0}^{\infty}\left(\prod_{k=0}^{n} \frac{a+b+k}{a+k+1}\right) q^{n+1}\right].$$

The summation is terminated whenever the summand becomes smaller than 10^{-10}. We compute Γa (Γ of any real number a) using logarithms and the formula $\Gamma a = (a-1)(a-2)\ldots(1+f)(f!)$, where f is the fractional part of the number a, and $f!$ is computed using formula 6.1.36 of [7]. The values of p, thus computed, are presented in table 1. The maximum absolute error in these computations is less than 10^{-6}. Here p is the probability of observing such a large, or larger, deviation at the Tay-Sachs disease locus.

If there are N such loci (rare autosomal traits with small equilibrium gene frequencies around $\bar{Q}$) segregating independently, we might wish to know the probability of observing such large deviations for at least one of the N loci. Large deviation being a rare event, we might as well assume a Poisson distribution and approximate the probability of the above event by $P_N(1) = 1 - e^{-Np}$. There is evidence [8] to believe that N is at least 500, and may even be 1,000 or 2,000. Assuming independent segregation of all the loci, we have computed $P_N(1)$ for $N = 500$ and 2,000. These values (table 1) are too large to reject the possibility of such large deviations for at least one locus in a given population. Presumably there are several loci for which such large deviations occur. We now turn to the probability, $P_N(k)$, of observing large deviations in the same population for at least k of the N loci, given by

$$P_N(k) = 1 - \sum_{r=0}^{k-1} e^{-Np}(Np)^r/r!.$$

As an example we present values of $P_N(5)$ for $N = 500$ and 2,000 in table 1. All entries of table 1 suggest that these events of large deviations for one or a few loci may be rare, but not so rare as to argue strongly against genetic drift.

The same methodology can be applied to any disease, for example, to cystic fibrosis. Wright and Morton [9] have reported a large gene frequency of .01616 among Caucasians in Hawaii, compared with .00331 for non-Caucasians. Supposing the latter to be the equilibrium value, and so with $\bar{Q} = .00331$ and $q = .01616$, we can compute all the entries of table 1 for cystic fibrosis also. The C, p, $P_N(1)$, and $P_N(5)$ are computed for this disease for $F = .0009$ (which was estimated from outcrossing effects in Hawaii as the inbreeding coefficient measuring differentiation of major racial groups for rare genes [10]) and for $F = .005$. These values are given in the last two rows of table 1. Here, too, the maximum absolute error is less than 10^{-6}. It is thus clear that chance alone could very well have caused such a large deviation for the cystic fibrosis locus also. Hence, it appears that there are at least two loci where large deviations have been observed (in different populations), and the magnitudes of $P_N(5)$ lead us to expect such observations at more loci in the future, in the absence of local selective advantage.

Finally, table 2 incorporates values of p for several combinations of $\bar{Q}$ (mean),

TABLE 2

VALUES OF p FOR VARIOUS COMBINATIONS OF $\bar{Q}$, q, AND F

$\bar{Q}$	q	F (Inbreeding)					
		.001	.003	.005	.007	.01	.03
.0010	.0100	.0000	.0051	.0130	.0189	.0243	.0279
	.0125	.0000	.0019	.0068	.0115	.0165	.0230
	.0150	.0000	.0008	.0037	.0071	.0114	.0192
.0015	.0100	.0002	.0098	.0220	.0306	.0380	.0420
	.0125	.0000	.0038	.0118	.0188	.0260	.0346
	.0150	.0000	.0015	.0065	.0118	.0181	.0290
	.0175	.0000	.0006	.0036	.0075	.0128	.0245
	.0200	.0000	.0002	.0020	.0048	.0091	.0209
.0020	.0100	.0005	.0162	.0329	.0437	.0527	.0562
	.0125	.0000	.0066	.0180	.0272	.0364	.0464
	.0150	.0000	.0027	.0100	.0172	.0255	.0389
	.0175	.0000	.0011	.0056	.0110	.0182	.0330
	.0200	.0000	.0005	.0031	.0071	.0130	.0282
.0025	.0150	.0000	.0044	.0142	.0235	.0336	.0489
	.0175	.0000	.0018	.0081	.0152	.0240	.0415
	.0200	.0000	.0008	.0046	.0099	.0173	.0356
.0030	.0175	.0000	.0028	.0111	.0199	.0304	.0503
	.0200	.0000	.0012	.0064	.0131	.0221	.0431
	.0225	.0000	.0005	.0037	.0086	.0161	.0372

NOTE.—See text for explanation.

q (observed), and F (inbreeding). The maximum absolute error is less than 10^{-6}. We hope that this table will be of some use in future investigations.

All these computations are incorporated in the computer program DRG, which is written in FORTRAN for the CDC 3100 [11]. The input specifies the values of $\bar{Q}$, q, and F.

It should be pointed out that these results are valid only to the extent that the gene frequency distribution can be approximated by a β function, which can be derived on the hypothesis of a linear systematic pressure on island populations [12]. It is our belief that this β approximation does reasonably well for extreme tail probabilities even if the true distribution deviates considerably from β.

A further test of drift is provided by the physiological independence of rare genes which reach high frequency in a particular population. Thus the simultaneous elevation of two ganglion lipidoses (Tay-Sachs and Niemann-Pick) in Ashkenazis, or of more than one gene for cystic fibrosis in Caucasians, provides an argument against drift, but the evidence is difficult to evaluate statistically and favors the Scotch verdict of "not proven."

SUMMARY

A gene rare in a species is sometimes found to have an appreciable frequency in a restricted population, suggesting either genetic drift, founder effect, or local selective advantage. Here we show that large deviations from the mean gene frequency are not improbable and they could very well occur due to chance alone. Methodology is illustrated for Tay-Sachs disease and cystic fibrosis.

REFERENCES

1. CHASE GA, MCKUSICK VA: Founder effect in Tay-Sachs disease. *Am J Hum Genet* 24:339–340, 1972
2. MYRIANTHOPOULOS NC, NAYLOR AF, ARONSON SM: Founder effect in Tay-Sachs disease unlikely. *Am J Hum Genet* 24:341–342, 1972
3. MYRIANTHOPOULOS NC, ARONSON SM: Population dynamics of Tay-Sachs disease. I. Reproductive fitness and selection. *Am J Hum Genet* 18:313–327, 1966
4. SHAW RF, SMITH AP: Is Tay-Sachs disease increasing? *Nature* (Lond) 224:1214–1215, 1969
5. MYRIANTHOPOULOS NC, NAYLOR AF, ARONSON SM: Tay-Sachs disease is probably not increasing. *Nature* (Lond) 227:609, 1970
6. RAO CR: *Linear Statistical Inference and Its Applications.* New York, Wiley, 1965
7. ABRAMOWITZ M, STEGUN IA (eds): *Handbook of Mathematical Functions.* New York, Dover, 1965
8. MCKUSICK VA: *Mendelian Inheritance in Man.* Baltimore, Johns Hopkins Press, 1966
9. WRIGHT SW, MORTON NE: Genetic studies on cystic fibrosis in Hawaii. *Am J Hum Genet* 20:157–169, 1968
10. MORTON NE, CHUNG CS, MI MP: *Genetics of Interracial Crosses in Hawaii,* Monographs in Human Genetics, vol 3. Basel, S. Karger, 1967
11. RAO DC: DRG, in *Genetic Structure of Populations,* edited by MORTON NE, Honolulu, Univ. Hawaii Press, 1973
12. WRIGHT S: Evolution in the Mendelian populations. *Genetics* 16:97–159, 1930