Am. J. Hum. Genet. 46:452–458, 1990

The Estimation of Selection Coefficients in Afrikaners: Huntington Disease, Porphyria Variegata, and Lipoid Proteinosis

O. Colin Stine and Kirby D. Smith

Division of Medical Genetics, Johns Hopkins Medical Institutions, Baltimore, MD

Summary

The effects of mutation, migration, random drift, and selection on the change in frequency of the alleles associated with Huntington disease, porphyria variegata, and lipoid proteinosis have been assessed in the Afrikaner population of South Africa. Although admixture cannot be completely discounted, it was possible to exclude migration and new mutation as major sources of changes in the frequency of these alleles by limiting analyses to pedigrees descendant from founding families. Calculations which overestimated the possible effect of random drift demonstrated that drift did not account for the observed changes in gene frequencies. Therefore these changes must have been caused by natural selection, and a coefficient of selection was estimated for each trait. For the rare, dominant, deleterious allele associated with Huntington disease, the coefficient of selection was estimated to be .34, indicating that this allele has a selective disadvantage, contrary to some recent studies. For the presumed dominant and probably deleterious allele associated with porphyria variegata, the coefficient of selection lies between .07 and .02. The coefficient of selection for the rare, clinically recessive allele associated with lipoid proteinosis was estimated to be .07. Calculations based on a model system indicate that the observed decrease in allele frequency cannot be explained solely on the basis of selection against the homozygote. Thus, this may be an example of a pleiotropic gene which has a dominant effect in terms of selection even though its known clinical effect is recessive.

Introduction

The effect of selection on gene frequency has rarely been measured in human populations. In the few cases where it has been estimated, independent determinations are often in conflict or give unanticipated results. For example, the allele associated with Huntington disease was measured to be selectively deleterious in a Michigan population (Reed and Neel 1959) but, surprisingly, was determined to be selectively neutral in a Queensland population (Wallace and Parker 1973) and selectively advantageous in both a Minnesota (Marx 1973) and a Welsh (Walker et al. 1983) population. Alleles at loci associated with the genetic diseases porphyria variegata and lipoid proteinosis have been postulated to be selectively neutral, despite their effect on phenotype, because of the high incidence of these diseases in South Africa (Dean 1971; Botha and Beighton 1983*b*). More recently, the high incidence has been attributed to founder effect (Diamond and Rotter 1987). In the present paper, estimates of the coefficients of selection for the alleles associated with each of these traits have been made in the well-defined Afrikaner population of South Africa.

Directional selection alters gene frequency from one generation to the next within a population. Before selection can be estimated from an observed change in gene frequency between generations, three alternative explanations—migration, mutation, and random drift—must be examined. Pedigree analysis in the well-defined Afrikaner population would allow new immigrants and new mutations to be easily recognized. The effect of random drift was determined by estimating its maximum possible effect given the initial population size and gene frequency and then comparing this value with

Received April 4, 1989; revision received August 24, 1989.

Address for correspondence and reprints: O. Colin Stine, Division of Medical Genetics, 933 Traylor Building, Johns Hopkins Medical Institutions, 720 Rutland Avenue, Baltimore, MD 21205.

0002-9297/90/4603-0007$02.00

the observed generational change in gene frequency. In each case the observed change in gene frequency was larger than the estimated maximum effect of random drift. Thus selection must have occurred, and a coefficient of selection could be estimated. The coefficient of selection measures the relative lack of offspring produced by an individual with a given genotype. For example, a coefficient of selection of .2 would mean that affected individuals produce 20% fewer offspring than does an average individual in that population.

The Afrikaner population of South Africa was used to estimate the effects of selection on the alleles associated with Huntington disease, porphyria variegata, and lipoid proteinosis. The Afrikaners are a subset of white South Africans whose ancestry has been traced to a small group of Dutch, German, and Huguenot settlers who founded the Cape Colony in the late 17th and early 18th centuries. There have been 12–14 generations since the founding, and it has been estimated that currently there are 2.5×10^6 Afrikaners in South Africa (Botha and Beighton 1983*b*). There are conflicting estimates of the size of the founding population. Dean (1972) states that there were 20 original family names—and thus presumably 20 separate families in the founding Afrikaner population—while Botha and Beighton (1983*b*) state that there were 50 original free burghers and their wives. Botha and Beighton (1983*b*) refer to a list of immigrants which shows that half of the most common surnames among the Afrikaners arrived before 1691 and that the other half arrived before 1717. The later arrivals could be treated either as immigrants to the second generation or as members of the first generation if the population had discrete generations. However, this population has overlapping generations, and the status of the later arrivals is uncertain. Although the number of original settlers is not clear, an initial population of 100 will be used to calculate the effect of random drift, the observed change in gene frequencies, and the coefficients of selection. This estimate is more inclusive and is consistent with the 1691 and 1701 censuses, which count all whites of all ages (see Botha 1972; Botha and Beighton 1983*a*).

The analysis was limited to Afrikaners in an attempt to limit the effect of migration on the change in gene frequencies. If white South Africans were considered as the current population, the change in gene frequencies due to English immigrants and their progeny would have to be calculated. The calculations are more straightforward if these migrations are excluded.

Some small migrations—such as the immigration of small groups absorbed by the Afrikaners; the admixture by Bantu, colored, and English; and the emigration of Afrikaners both to other nations (notably Zimbabwe and Zambia) and to the colored population of South Africa (Dean 1971)—have affected the Afrikaner population. While available data do not allow accurate estimation of the magnitude of these migrations, it has been estimated from blood group markers that the Afrikaner population has approximately 8% Bantu admixture (Botha 1972). In addition, we show here (see below) that even admixtures of up to 50% have only a small effect on our calculations. Thus, the effect of migration can be ignored in our calculation of the coefficients of selection.

The presence of affected individuals in the population because of immigration or new mutation was eliminated by studying traits which could be traced by pedigree analysis, to a carrier among the founding colonists. For each of the diseases to be discussed, the pedigrees of affected individuals from different families could be traced to single families among the founders of the Cape Colony (Dean 1971; Hayden 1981; Botha and Beighton 1983*b*). This is consistent with the hypothesis that these alleles were present in the founders. Given that these mutations are rare elsewhere in the world, it is unnecessary to postulate a new mutation or a new immigrant as the origin of these alleles in any affected individual. Even if some of the affected individuals did result from new mutations, the calculations would only slightly underestimate the deleterious effects of these alleles.

For determination of the effect of selection on gene frequency, the critical comparison is between the observed generational change in gene frequency and the magnitude of the possible effect due to random drift. If the estimated magnitude of random drift cannot account for the observed change, then natural selection has occurred. Although the effect of random drift on gene frequencies in the Afrikaner population cannot be determined exactly, its maximum possible effect was estimated from the change in gene frequency between successive generations, according to the equation $pq/2N$, where p and q are the frequencies of alternative alleles and N is the population size (Falconer 1976, p. 51). This expression is maximized when N is smallest. In the rapidly expanding Afrikaner lineage (Botha and Beighton 1983*a*), this occurred in the first generation.

Coefficients of selection were estimated using the method of Clarke and Murray (1962). Their model assumes (1) a large, Mendelian population in which there is no mutation or migration, (2) an allele which is dominant and deleterious, and (3) discrete generations. Although Afrikaners have overlapping generations, Hal-

dane (1926) has determined that this does not significantly affect the outcome of these types of calculations. A consequence of this model is that for rare alleles the coefficient of selection is primarily derived from affected heterozygotes, because of the virtual absence of homozygotes.

According to this model (Clarke and Murray 1962), the wild-type homozygote occurs at a frequency of q^2 and has a relative fitness of 1, the heterozygote occurs at a frequency of $2pq$ and has a relative fitness of 1 minus the selection coefficient (S), and the affected homozygote occurs at a frequency of p^2 and also has a relative fitness of $1 - S$. Given these assumptions, the change in q from the initial to the next generation is $[Sq_o^2(1-q_o)]/(1-Sq_o^2)$, where q_o is the gene frequency in the initial generation (see Falconer 1976, table 2.1). When treated as a differential equation, so that n generations can be considered, the solution after integration can be expressed in terms of the initial and final gene frequencies q_o and q_n and the number of generations n:

$$S = \frac{\log_e\left[\frac{q_o\,(1-q_n)}{q_n\,(1-q_o)}\right] + \frac{1}{q_n} - \frac{1}{q_o}}{t + \log_e\left(\frac{1-q_n}{1-q_o}\right)}.$$

The variance of the selection coefficient $V(S)$ is also dependent on the initial and final sample sizes N_o and N_n:

$$V(S) = \frac{\left(\frac{1}{q_n} + \frac{1}{1-q_n} + \frac{1}{q_n^2}\right)^2 \frac{q_n(1-q_n)}{2N_n} + \left(\frac{1}{q_o} + \frac{1}{1-q_o} + \frac{1}{q_o^2}\right)^2 \frac{q_o(1-q_o)}{2N_o}}{\left[t + \log_e\left(\frac{1-q_n}{1-q_o}\right)\right]^2}.$$

Huntington Disease

A recent survey identified 210 Afrikaners affected with Huntington disease who were descendants of an original settler who was a presumed heterozygote (Hayden 1981). Thus, the gene frequency in the original population was 5×10^{-3} (1/200). When the expression $pq/2N$ is applied, the estimated magnitude of change due to random drift from the first to the second generation is 4.98×10^{-5}. Despite the very small amount of change that can be attributed to random drift in the first generation, it is the maximum per generation because the population was expanding in size. Thus, using the magnitude of change associated with the first generation as an estimate for each of the succeeding 14 generations will overestimate the accumulated effect of random drift between the initial and current generations. In the extremely unlikely event that the gene frequency persistently drifted in the same direction in each of the 14 generations (2^{-14}, or 1/16,384), the maximum change in gene frequency that could have occurred would be 7×10^{-4}. The current gene frequency estimated from the 210 affected individuals (Hayden 1981) identified among the 2.5×10^6 Afrikaners descendant from the original colonists (Botha and Beighton 1983*a*) is 4.2×10^{-5}. Thus, the current gene frequency is approximately two orders of magnitude less than the initial gene frequency. The observed change in gene frequency, 4.9×10^{-3}, is much too large to be explained by random drift alone. Therefore, selection must be operating to reduce the frequency of the allele associated with Huntington disease.

If the original frequency plus or minus the maximum accumulated effect due to random drift was found in the current population, there would currently be 25,000 ± 3,500 affected individuals instead of the 210 observed affected individuals. This substantial decrease must be attributed primarily to natural selection, consistent with the conclusion of Reed and Neel (1959) that the allele associated with Huntington disease is deleterious.

The calculated coefficient of selection against individuals in the Afrikaner lineage who have the allele for Huntington disease is .3419 (table 1). This estimate of selective disadvantage is twice that estimated by Reed and Neel (1959) and may reflect differences either in the populations studied or in the nature of the mutation. Juvenile onset occurs in the Afrikaners (Hayden 1981) but not in the Michigan population studied by Reed and Neel (1959). In addition, onset occurs before the age of 30 years more frequently in Afrikaners (33.3%) than in Michiganders (26.3%) (Hayden 1981). Earlier onset could affect the coefficient of selection because the affected Afrikaners would be expected to be reproductively active for a shorter period of time—i.e., to have a smaller reproductive value (Fisher 1958)—and to provide less maternal care (see Reed and Neel 1959) than the Michiganders.

The estimate of current gene frequency in the Afrikaners would also be affected by any admixture from other populations. However, the effect on these calculations would be small. For example, if 50% of the current Afrikaner population is the result of admixture,

Table 1

A Numerical Summary

	Huntington's Disease	Porphyria	Lipoid Proteinosis
Initial population size	100	100	100
Initial frequency:			
Mutant	.005	.005	.01
Wild type	.995	.995	.99
Changes due to random drift:			
1 Generation	4.98×10^{-5}	4.98×10^{-5}	9.9×10^{-5}
n Generations	7.0×10^{-4}	6.0×10^{-4}	1.39×10^{-3}
Recent gene frequency	4.2×10^{-5}	Maximum .00225, minimum .00394	.0036
Observed change (mutant)	.0049	Maximum .00275, minimum .00116	.0064
Explained by random drift	14%	Maximum 22%, minimum 51%	22%
Coefficient of selection	.3419	Maximum .067, minimum .020	.0742
Variance	.0052	.0073	.0026

the size of the population descendant from the founders would be 1.25×10^6, but the coefficient of selection would only be reduced to .2924. Clearly, migration can only account for a fraction of the observed change.

In the Australian study (Wallace and Parker 1973), the allele associated with Huntington disease was estimated to be selectively neutral ($S = 0$). The selective differences between the Australians and the Americans and Afrikaners could be due to (1) later onset in Australians (only 17.3% are affected before age 30 years), (2) the presence of a remarkably large Australian family which skews the data (for a previous example, see Reed and Palm 1951, and for discussion, see Reed and Neel 1959), and (3) the fact that the Australian estimate does not include early childhood mortality, which in Michigan was higher among the offspring of female choreics than among offspring of normal females (Reed and Neel 1959).

Surprisingly, two other studies, one in Minnesota and one in South Wales, which estimated the coefficient of selection (1 − fitness) for individuals with Huntington disease found the allele to be advantageous. In the first, Marx (1973) noted that many affected individuals were immigrants and that the immigrants as a group tended to have larger families than did the general population in Minnesota. Since the general population was used as the control group, it is likely that the estimated coefficient of selection contains a cultural component and thus is not an accurate estimate of the selective value of the allele associated with this disorder.

The second study (Walker et al. 1983) used data from an earlier survey (Walker et al. 1981) that had an ascertainment method which might inflate fitness estimates in two ways (both acknowledged by the authors in 1981). Their method was to identify propositi and then intensively search among relatives for other affected individuals. Only 30% of the affected individuals were propositi, an observation consistent with two possible biases. First, families with only a few affected individuals may have been undetected. Second, large families may be overrepresented. Indeed, Walker et al. (1981) remarked that some affected individuals in small families were diagnosed after the close of the study and that in one of the two counties there are two very large kindreds which skew the data (see above).

Porphyria Variegata

A recent survey identified 29 individuals with porphyria variegata among 6,458 adult patients at two Port Elizabeth hospitals (Dean 1971): a gene frequency of 2.25×10^{-3}. All of the affected individuals are descendants of a single original founder who was a presumed heterozygote giving rise to both affected and nonaffected lineages (Dean 1971). The calculated gene frequency is a minimum because it is unlikely that all of the hospital patients are Afrikaners descendant from the founding population. Nationally, 57% (Botha and Beighton 1983*a*) of white South Africans are Afrikaners. Thus, the expected number of patients of Afrikaner

descent can be estimated to be 3,681. Since Dean (1971) states that the proportion of Afrikaners is higher in the Port Elizabeth area than in South Africa as a whole, this is likely to be an underestimate. Thus, the range of estimated current gene frequencies in the sample population is 2.25×10^{-3} (N = 6,458)–3.94×10^{-3} (N = 3,681).

Since all Afrikaner porphyrics are derived from a single heterozygous individual, the initial gene frequency is 5×10^{-3}. The estimated magnitude of change due to random drift in one generation is 4.98×10^{-5}, and the maximum possible change in gene frequency over 12 generations is 6×10^{-4}. If the original frequency plus or minus the maximum accumulated effect due to random drift were found in the sample population, there would be 64 ± 4 (N = 6,458) or 37 ± 2 (N = 3,681) instead of the 29 observed affected individuals.

The observed change in gene frequency, 2.75×10^{-3}–1.16×10^{-3}, is too large to attribute to drift. Thus, the allele associated with porphyria variegata is selectively deleterious, and a range of coefficients of selection may be calculated from the previously described model (Clarke and Murray 1962). In keeping with this model, the allele was assumed to be dominant both clinically and selectively. As presented in table 1, coefficients of selection were estimated for the two extreme estimates of the sample gene frequency. Based on these values, the coefficient of selection ranges from .067 to .020.

It is possible that a significant proportion of this selective disadvantage may be ascribed to the nearly lethal synergism between porphyria and barbiturates which affected this population during the first half of the 20th century. While it is not possible to quantify the effect of barbiturates on the allele frequency of porphyria because the number of affected individuals is not known, it may well be that selection against this allele is significantly reduced in the absence of barbiturates. In addition, for this trait some or all of the changes in gene frequency may be accounted for without recourse to selection. For example, if the minimum estimated sample size is correct (unlikely; see above) and if 20% of the current Afrikaner population is the result of admixture, then the observed change in gene frequency can be accounted for by migration. This would imply that the allele associated with porphyria might be selectively neutral or even advantageous. However, if the maximum estimated sample size is correct, even 50% admixture cannot account for the change in allele frequency. Thus, although there is considerable uncertainty as to the value of the coefficient of selection, porphyria is probably deleterious. This conclusion is consistent with calculations by Dean (1971). He estimated that there should be 32–36 affected individuals in the Port Elizabeth survey. He derived these values from the number of individuals surveyed who had the name van Rooyen; porphyria is commonly called van Rooyen disease. Thus, if these estimates are considered the expected values in the absence of selection, the 29 observed porphyria patients reflect the selective disadvantage.

A direct test to determine the deleteriousness of porphyria variegata is possible if appropriate Afrikaner populations can be identified and examined. For example, the first trekkers into an area could be used as the initial population and a random sample of their living descendants, tested for porphyria variegata, could be used to determine the current gene frequency.

Lipoid Proteinosis

There are 32 identified individuals affected with lipoid proteinosis among living Afrikaners. Pedigree analysis indicates that all are descendants of two (related) original colonists (Botha and Beighton 1983*b*). Although it is difficult to rule out new mutations to the clinically recessive allele associated with lipoid proteinosis, the extreme rarity of this trait in the rest of the world (one-third of reported cases are in South Africa; Botha and Beighton 1983*b*) and the consistent pedigrees suggest that new mutations have not contributed significantly to the current gene frequency. (If a new mutation has occurred, its effect on the calculations is to strengthen our conclusions.) As the clinically recognized trait is recessive, each affected individual has two copies of the allele and there are expected to be many heterozygous carriers. If the population is assumed to be in Hardy-Weinberg equilibrium, the estimated current gene frequency is 3.6×10^{-3}. Given that the affected individuals are descendants of two original colonists, the gene frequency in the initial population is .01. The estimated maximum of change due to random drift in a single generation, calculated from the formula $pq/2N$, is 9.9×10^{-5}. Over 14 generations, the maximum effect is 1.39×10^{-3}. Thus, at a maximum, random drift could account for less than one-fourth of the observed change in gene frequency (see table 1). If the original frequency plus or minus the maximum accumulated effect due to random drift had been found in the current population, there would currently be 250 ± 5 instead of the 32 observed affected individuals. This substantial decrease must be attributed primarily to natural selection.

The formula for calculating the selection coefficient in the case of a recessive trait is modified from the model for a dominant trait (see Clarke and Murray 1962). Since lipoid proteinosis is an autosomal recessive and since heterozygotes appear normal, selection against heterozygotes is not expected. Therefore, the fitness of both the wild-type homozygote and the heterozygote will be 1, and their frequencies will be p^2 and $2pq$, respectively. The frequency of the mutant homozygote will be q^2 and their fitness will be $1 - S$. If the initial population size and gene frequencies for lipoid proteinosis discussed above are used and if that selection is assumed to be only against the homozygous recessive, the calculated coefficient of selection is always greater than 1, even if allowance is made for the maximum possible effect of random drift. By definition, this is impossible, since a person can only die leaving no offspring once. Thus, the change in gene frequency cannot be attributed to selection solely against the recessive homozygote, implying that the heterozygote is also selectively disadvantageous. Thus, the mutant allele associated with lipoid proteinosis, although recessive to the wild-type allele in terms of the clinically recognized phenotype, must be dominant in terms of fitness. Indeed, Fisher (1958) predicted that alleles with these particular phenotypic relationships would be common.

If the mutant allele is assumed to be dominant, the coefficient of selection can be calculated employing the previously described assumptions and the model for a dominant allele. The coefficient of selection is .0743 and is appropriately attributed to the heterozygotes, since they outnumber the affected homozygotes by at least 200 to 1 in every generation. Thus even if the affected homozygote were lethal ($S = 1$), the calculated coefficient would change very little. In addition, if 50% of the current Afrikaner population were the result of recent admixture, the coefficient of selection would be reduced to .0493. It is tempting to speculate that, if natural selection adversely affects the heterozygote, there may be additional, as yet unrecognized, clinical phenotypes associated with this allele.

Conclusion

The coefficients of selection estimated in the present study are small, reflecting a slight but significant reduction each generation in the number of offspring produced by mutant heterozygotes compared with normal homozygotes. Such small differences can only be measured if there is a very large sample size (e.g., see Reed and Neel 1959) or if a population can be studied over several generations as with the Afrikaners. In order to measure the change in gene frequency over several generations, it is essential that the population have a known pedigree. The ability to trace the lineage of affected individuals minimizes ascertainment problems as well as allowing the effect of mutation and migration to be determined. Once these two forces are accounted for, the effect of random drift on gene frequency can be estimated. When the observed frequency change cannot be explained by any combination of these forces, it must be attributed to natural selection. It is important to recognize that even though natural selection has proportionally decreased the frequency of the alleles associated with Huntington disease, porphyria variegata, and lipoid proteinosis, the enormous expansion of population size has permitted the survival of a large number of mutant phenotypes producing the perceived high incidence of these diseases among Afrikaners.

Acknowledgment

We appreciate the helpful comments of our reviewers and the editor.

References

Botha MC (1972) Blood group gene frequencies: an indication of the genetic constitution of population samples in Cape Town. S Afr Med J [Suppl 1]: 1–27

Botha MC, Beighton P (1983*a*) Inherited disorders in the Afrikaner population of southern Africa. Part I. Historical and demographic background, cardiovascular, neurological, metabolic and intestinal conditions. S Afr Med J 64:609–612

——— (1983*b*) Inherited disorders in the Afrikaner population of Southern Africa. Part II. Skeletal, dermal and haematological conditions; the Afrikaners of Gamkaskloof; demographic considerations. S Afr Med J 64:664–667

Clarke B, Murray J (1962) Changes of gene frequency in *Cepaea nemoralis* (L): the estimation of selective values. Heredity 17:467–476

Dean G (1971) The porphyrias. J. B. Lippincott, Philadelphia

Diamond JM, Rotter JI (1987) Observing founder effect in human evolution. Nature 329:105–106

Falconer DS (1976) Introduction to quantitative genetics. Ronald, New York

Fisher RA (1958) The genetical theory of natural selection. Dover, New York

Haldane JBS (1926) A mathematical theory of natural and artificial selection, IV. Proc Camb Phil Soc 23:607–615

Hayden MR (1981) Huntington's chorea. Springer, New York

Marx RN (1973) Huntington's chorea in Minnesota. Adv Neurol 1:237–243

Reed TE, Neel JV (1959) Huntington's chorea in Michigan. II. Selection and mutation. Am J Hum Genet 11:107–136

Reed SC, Palm JD (1951) Social fitness versus reproductive fitness. Science 113:294–296

Walker DA, Harper PS, Newcombe RG, Davies K (1983) Huntington's chorea in South Wales: mutation, fertility and genetic fitness. J Med Genet 20:12–17

Walker DA, Harper PS, Wells CEC, Tyler A, Davies K, Newcombe RG (1981) Huntington's chorea in South Wales: a genetic and epidemiological study. Clin Gene 19:213–221

Wallace DC, Parker N (1973) Huntington's chorea in Queensland: the most recent story. Adv Neurol 1:223–236