THE RELATIVE IMPORTANCE OF PRINCIPAL AND MODIFYING GENES IN DETERMINING SOME HUMAN DISEASES

By J. B. S. HALDANE, F.R.S.

(With One Text-figure)

A VAST amount of rather speculative theories have been produced regarding natural selection in man. It is probable that in modern civilized communities reproductive selection, that is to say, selection on a basis of fertility, is more important than strictly Darwinian selection on the basis of survival. But the opposite was probably the case through most of man's career. Moreover, what little exact genetical knowledge we possess relates wholly to selection on the basis of survival.

A number of genes are known which diminish the expectation of human life, but yet allow long enough life to permit of reproduction by many individuals. Bell (1939) has shown that in a number of such cases the period between the onset of the disease and death is almost independent of the age of onset. Thus the age of onset is a valuable index of the selective disadvantage of the gene.

With the discovery of modifying genes there has been a tendency in some quarters to postulate their existence and importance on a very wide scale, while other writers have stressed the importance of environment. Among a group of persons affected with a given disease, whose genetical basis is roughly known, we can examine the relative importance in determining the age of onset, of:

(1) Differences in the main gene itself. That is to say, there may be two or more different main genes, each giving a characteristic mean age of onset, with some variation round it.

(2) Differences in modifying genes, which, while not causing the disease, may favour early or late onset if a main gene is present.

(3) Differences in environment, which again may favour early or late onset.

The recent work of Bell (1934, 1935, 1939) enables us, for the first time, to answer this fundamental question in certain cases. Bell takes a large group of patients, and finds the correlation between age of onset in pairs of sibs (and sometimes also in parents and offspring). Let us see what we should expect on various extreme hypotheses. It is not

suggested that any one is ever exactly true, but one or other may prove a good approximation to the truth in a particular case.

If the age of onset were determined wholly by the main gene, and if there were only one main gene, we should expect to find the age of onset exactly the same in all cases, say 7 years. But if there were several different genes we should expect to find several different ages of onset, say one group at 4 years, another at 12, and another at 30. If there were several different recessive allelomorphs, some ages of onset would characterize homozygotes, and these would almost invariably be found where the parents were related. Where they were unrelated we should sometimes find heterozygous "compound" recessive types, probably with intermediate ages of onset. In every case of dominance parent and offspring would be affected by the same gene, and have the same age of onset. Thus the correlation would be unity. Two sibs would always have the same pair of recessive main genes except where one parent was affected, and happened to be a heterozygous compound. Thus the correlation between sibs would be very close to unity.

Next suppose that only one main gene were present in the population, but that there were also a number of modifying genes which affected the age of onset. For example, glaucoma is due to abnormally high pressure within the eye. We might expect the gene or genes which are responsible for high arterial pressure to accelerate its onset. If there were only one main gene and a considerable number of modifiers we should expect to find the same situation as with characters such as stature, which appears to be controlled by many genes. The correlation coefficients would be about 0·5, probably a little higher for sibs, and a little lower for parents and offspring. In the case of a dominant gene, the normal parent would, on the average, be responsible for just as much modification as the affected parent, though we should in general have no clue as to the nature of the modifiers to be found in a normal individual.

If there were one main gene and no modifiers, but environment played a large part in determining the age of onset, we should expect to find some correlation, but its value would be low and uncertain. It would be much higher between sibs than between parent and offspring. For sibs are generally brought up in a similar environment, but the family environment may change a great deal within a generation. It would also be higher in a disease manifesting itself in childhood than later in life, since sibs are brought up in the same home and may separate later on. If environment were important in determining the age of onset we might

expect to find this fact recognized, and there might be large occupational and sex differences, as there are with cancer, which undoubtedly has a genetical basis in many instances. The influence of the environment is recognized in the case of the congenital photosensitivities, such as haematoporphyria and xeroderma pigmentosum. If there were no genetical basis for correlation, we should expect to find coefficients well below 0·5 in most cases.

Bell's findings are summarized in Table 1. In no case except the first is there any marked sex difference in frequency or in age of onset.

Table 1. *Correlation coefficients of ages of onset (after Bell)*

		Pairs of siblings		Parent and offspring	
Disease	Type	n	r	n	r
Optic atrophy (males only)	S.L. Rec.	812	0·510	—	—
Glaucoma	Dom.	256	0·897	113	0·813
Huntington's chorea	Dom.	442	0·465	153	0·593
Peroneal atrophy	Dom.	164	0·803	0·8	0·764
	Rec.*	108	0·840	—	—
Friedreich's ataxia	Dom.	144	0·925	—	—
	Rec.	500	0·694	—	—
Spastic ataxia	Dom.	198	0·812	—	—
	Rec.	164	0·845	—	—
Spastic paraplegia	Dom.	154	0·884	—	—
	Rec.	218	0·852	—	—
Grouped ataxias and paraplegia	Dom.	—	—	135	0·743

* Includes a few sex-linked cases.

The last line refers to grouped cases of dominant Friedreich's ataxia, spastic ataxia, and spastic paraplegia, which in Bell's opinion, though not in my own, may be manifestations of the same fundamental inherited abnormality. The precise value to be attached to the significance of these coefficients is rather uncertain, because except in the case of Huntington's chorea, the distribution of ages of onset is highly asymmetrical, and the theory of normal correlation does not apply. However, there is little doubt that in most cases the values are significantly above 0·5. Thus in the case of recessive Friedreich's ataxia, the value of Fisher's (1938) transformed coefficient z is $0{\cdot}856 \pm 0{\cdot}045$. This differs from $z = 0{\cdot}549$, the value corresponding with $r = 0{\cdot}5$, by nearly seven times its standard error. In the case of dominant peroneal atrophy, when parents and offspring are compared, $z = 0{\cdot}996 \pm 0{\cdot}109$, and the difference from 0·549 is over four times its standard error. On the other hand, the difference between the transformed fraternal and parental correlations for Huntington's chorea is about 1·9 times its standard error, and therefore not quite significant.

Actually these coefficients are, if anything, underestimated. Bell points out that in the case of the ataxias and spastic paraplegia the age of onset is, on an average, earlier in younger than elder sibs, perhaps because the parents are on the look-out for signs of the disease. Thus the coefficient of correlation is slightly raised if the table is made asymmetrical, age of onset being correlated in elder against younger sibs.

She has kindly calculated the effect of this in the case of recessive Friedreich's ataxia, and finds that it is raised from 0·694 to 0·711. Further, she has separated the dominant and recessive forms of the same disease which cannot be distinguished clinically. The ages of onset in the two groups overlap, but it is much later on the average in the dominants. If this had not been done, the correlations would be much higher. Thus the coefficient for dominant and recessive spastic ataxia combined is 0·890.

The calculation of the correlations between age of onset in sibs may be regarded as an analysis of variance, the variance within sibships being small compared with that between sibships when r is high. Thus in the case of dominant Friedreich's ataxia, $0{\cdot}925^2$ or 85·6% of the variance is between sibships, only 14·4% being within them.

I think that there is no escape from the conclusion that several different main genes are concerned in the causation of the diseases other than optic atrophy and Huntington's chorea. On no other hypothesis could the correlation coefficients be so large. The argument is particularly clear in the case of parent-offspring correlation. When the age of onset of glaucoma in a parent is known we have eliminated a fraction $1-r^2$, or 0·661 of the variance in the ages of onset in the offspring. Suppose that both parents influence this age by contributing modifiers, this fraction would be expected to be about $\frac{1}{4}$, since half the modifiers from each parent are lost by segregation. Even if we neglect segregation, there is not enough variance left over for the normal parent's modifiers to produce unless we postulate assortative mating so intense as to give rise to a correlation coefficient of over 0·5 for the modifying genes in the parents.

If we could distinguish the different main genes we could reduce the correlation coefficients to more normal values. Thus by dividing up the genes responsible for spastic ataxia into a dominant and recessive class, Bell was able to reduce a coefficient of 0·890 to two of 0·845 and 0·812. If we divide the pedigrees of dominant spastic ataxia into two groups, in one of which the age of onset is always under 50, in the other always over, we obtain two correlation tables, with $n=162$, $r=0{\cdot}562$, and $n=36$,

$r = 0{\cdot}027$. It is not of course suggested that this represents a real genetical division, but it happens that in this particular table all sibships can be divided into those with ages of onset over or under 50.

In the same way we can divide her correlation table for dominant Friedreich's ataxia into three parts. She considered seventy-two pairs of sibs, so there were 144 entries in the table. In the forty cases where the age of onset in one sib was under 5, it was invariably so in the other. In the twelve cases where it was between 35 and 39 in one sib, it was also so in the other.

The argument for the existence of numerous genes is made all the more plausible because Bell has already shown that in the last four diseases of Table 1 there is a dominant and a recessive form. Besides these there is a sex-linked recessive form of peroneal atrophy, and in one pedigree (490) spastic paraplegia behaves as a sex-linked recessive. If then three different genes may be responsible for a group of cases which are indistinguishable clinically, why not five or six? The different genes may in some cases be allelomorphic. The demonstration of this must await linkage tests. If the argument of Haldane (1940*a*) is accepted, recessive spastic paraplegia is due to a group of partially sex-linked genes, and their allelomorphism is highly probable. The dominant form is due to autosomal genes.

Pedigree 464 (due to Fry) suggests an interesting possibility. It is shown in Fig. 1. The grandfather of the three affected children was healthy up to 65, when his gait became uncertain. At 86 he had marked

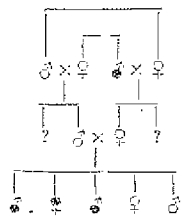

Fig. 1.

ataxia and other symptoms. He lived till 91. The grandchildren developed the disease at the age of about 10. As their parents were double first cousins we naturally suspect a recessive gene. It seems likely that the affected grandfather and his sister or her husband were both heterozygous for it, as were the parents of the affected sibs, but that it only showed in one heterozygote, and that very late in life. Examples can be quoted, particularly from *Drosophila*, where a gene generally behaves as a recessive, but occasionally shows up as a "weak" dominant.

Similarly, in case 499, a woman who had married her uncle had a spastic paretic gait at the age of 72, but except for one attack in middle life, could walk without sticks till 62. Three daughters developed symptoms at about 40, and were severely crippled at 50. Several other cases, for example 385, where there is no inbreeding, are nevertheless suggestive of the same possibility, namely, that some of the genes concerned are incompletely recessive.

No doubt, even in the cases of high correlation, modifying genes may play a part. Bell's correlation tables show a few striking outliers. Thus in the case of recessive Friedreich's ataxia there were three cases out of 250 where the age of onset in sibs differed by over 35 years. If these were omitted from the table, the coefficient of correlation would be raised from 0·694 to 0·802. There is a strong suggestion that about 1% of the population may carry modifying genes which markedly delay the onset of this disease if the main gene is present.

Dr Bell has kindly permitted me to use a table prepared by her of the differences in age of onset of the disease in pairs of sibs. Besides calculating the mean difference we may proceed as follows. If each family were large enough there would be a mean age of onset in affected members, and a distribution round this mean. We can readily determine the even cumulants of this distribution.

These cumulants are half those of the differences, if we count each difference both as positive and negative, and make Sheppard's correction. Thus if ν_2 and ν_4 are the mean values of the squares and fourth powers of the differences, we have, for the distribution about the family mean,

$$\sigma^2 = \tfrac{1}{2}\nu_2 - \tfrac{1}{12}, \quad \gamma_2 = \beta_2 - 3 = \frac{2\nu_4 - 6\nu_2^2 + \frac{1}{30}}{(\nu_2 - \frac{1}{6})^2} = \frac{2\,(\nu_4 - \nu_2 + \frac{1}{10})}{(\nu_2 - \frac{1}{6})^2} - 6.$$

The unit of grouping is taken as a year. Actually it is larger in a fraction of the cases. Thus the true value of σ is slightly smaller than that given, the true value of γ_2 slightly larger. However, as Sheppard's correction never amounts to 2%, the further correction to be made is small. The results of this calculation are shown in Table 2.

Table 2. *Differences of age of onset within sibships*

Disease	Type	Mean	S.D.	γ_2
Friedreich's ataxia	Dom.	2·61	3·10	+ 6·52
	Rec.	2·47	3·38	+39·6
Spastic ataxia	Dom.	6·03	6·04	+ 2·47
	Rec.	3·51	6·14	+21·2
Spastic paraplegia	Dom.	3·99	5·01	+12·5
	Rec.	3·01	5·01	+26·5

A somewhat more accurate value of γ_2 could be obtained from the complete data. The standard deviation round the family mean varies between 3 and 6 years, and the values of γ_2 are all positive, and are large in the case of recessives. This means that there are fewer small and more large deviations than in a normal distribution with the same standard deviation, particularly in the case of recessives. This is what we should expect if there were a few modifying genes with considerable effect, but in most families there was very little modification. It is, however, curious that modifiers for recessives seem to be more frequent or more effective than those for dominants.

There is some evidence for the existence of modifying genes in the case of Huntington's chorea. Bell found that of 385 affected males 60·3% had affected fathers, and of 336 affected females 51·5% had affected mothers. If we consider grandparents also we have the results of Table 3.

Table 3. *Huntington's chorea*

	Affected		Normal		
Affected ancestors	Males	Females	Males	Females	q
Mother and grandmother	36	47	29	24	0·559
Mother and grandfather	43	59	32	32	0·548
Father and grandmother	36	36	24	41	0·438
Father and grandfather	92	55	53	77	0·394

In this table normals are not included unless they appear to have reached the age of 30. q is the frequency of affected females plus normal males. These results are consistent with the theory (Haldane, 1936) that some of the genes which modify the age of onset are sex-limited in their effect. Thus if there are modifiers in a family increasing the age of onset in females there will be an excess of males among the affected and of females among the normal. This will account for the results tabulated. The differences between q values are not all significant, but their order is as expected.

DISCUSSION

This paper is of course of a preliminary nature. Data for ages of onset must exist for a number of diseases. Ages of death would in some cases be equally valuable or more so, since here there is no subjective element. Any other measurable character of the disease would be equally valuable, provided it is fairly definite. Thus in pedigrees of myopia we could use the strength in dioptres at some standard age, in dwarfism the

height, and so on. But a very variable character, such as the coagulation time in haemophilia, would be useless.

Differences in age of onset in man probably correspond with differences in penetrance in *Drosophila* or other insects. If we imagine all cases in a pedigree to be examined at the age of 25 only, late age of onset would appear as low penetrance. On Goldschmidt's theory genes act by determining the rates of processes. In man a morbid process may lead to manifest disease at any age. In an insect an abnormal developmental process will not produce any visible effect on the morphology unless it does so before the imago is fully formed. Thus a study of ages of onset can tell us a good deal more than a study of penetrance for equal numbers studied.

So far as concerns the question of evolution, Huntington's chorea seems to agree with the conditions postulated by Fisher (1931) in his theory of the evolution of dominance. Modifiers are presumably being selected which delay the age of onset. Perhaps Huntington's chorea was a disease of infancy in *Sinanthropus*. And if eugenic measures are not taken against it, it may be confined to old age in our remote descendants, in which case the main gene will spread, and homozygotes appear, so that it will be, in effect, a recessive disease. But in the case of the other conditions (except Leber's disease, which is either sex-linked or cytoplasmic) the main effect of selection will be to weed out those main genes which produce an early onset, particularly where the disease is dominant.

This is discussed in detail elsewhere (Haldane, 1940*b*). As pointed out by Bell, selection causes the dominant forms to be less severe than the recessives. Modifying genes may exist, but where the average difference in age of onset between sibs is only 2 or 3 years, they cannot be very important, and must be selected very slowly. A great deal more work will be required before we can judge whether the presence of modifiers in accordance with Fisher's theory is common or rare in the human species.

Summary

Bell's data on the age of onset of some human hereditary diseases are discussed. In glaucoma, peroneal atrophy, Friedreich's ataxia, spastic ataxia, and spastic paraplegia, the age of onset in all affected members of a pedigree is nearly the same, while different pedigrees differ widely. Thus a number of different main genes must be responsible for the clinically indistinguishable diseases in different families. In optic

atrophy and Huntington's chorea the differences of age of onset within a pedigree are nearly as large as those between different pedigrees. So the same main gene may be responsible for all cases, while modifying genes account for much of the difference in age of onset. The bearing of these facts on evolutionary theories is discussed.

REFERENCES

BELL, S. (1934). "Huntington's chorea." *Treas. Hum. Inher.* **4**, 1.

—— (1935). "On the peroneal type of progressive muscular atrophy." *Treas. Hum. Inher.* **4**, 69.

—— (1939). "On hereditary ataxia and spastic paraplegia." *Treas. Hum. Inher.* **4**, 141.

FISHER, R. A. (1931). "The evolution of dominance." *Biol. Rev.* **6**, 345.

—— (1938). *Statistical Methods for Research Workers*, 6th ed. Edinburgh: Oliver and Boyd.

HALDANE, J. B. S. (1936). "A search for incomplete sex-linkage in man." *Ann. Eugen., Lond.*, **7**, 28.

—— (1940*a*). "The partial sex-linkage of recessive spastic paraplegia." *J. Genet.* (in the Press).

—— (1940*b*). "The conflict between selection and mutation of harmful recessive genes." *Ann. Eugen., Lond.* (in the Press).