COMMENTARY ON *J. GENET.* CLASSIC

Haldane and the first estimates of the human mutation rate

(A commentary on J.B.S. Haldane 1935 *J. Genet*. **31**, 317–326; reprinted in this issue as a *J. Genet.* classic, pages 235–244)

MICHAEL W. NACHMAN*

Department of Ecology and Evolutionary Biology, P.O. Box 210088, Biosciences West Bldg, University of Arizona, Tucson, AZ 85721, USA

Mutation is the ultimate source of genetic novelty. As such, the rate at which new mutations arise is a central issue in genetics, with profound implications for both evolution and human health. Haldane's 1935 paper (reprinted in this issue, pages 235–244) was the first to provide a careful estimate of the mutation rate in humans.

In fact, previous studies had already provided estimates of mutation rates in other organisms (Muller 1928; Stadler 1932). The chief difficulty confronting experimentalists was the fact that mutations occur at a very low rate. Muller (1928) was able to measure mutation rates in *Drosophila melanogaster* using a scheme of balanced lethals. By working with flies, he was also able to breed many individuals and score large numbers of progeny easily. Stadler (1932), working with maize, was able to count millions of seeds to arrive at mutation rates mostly between 10^{-4} and 10^{-6} for eight different genes. These approaches were clearly not available in humans, and it was recognized that measures of the mutation rate had to come from other methods.

Haldane (1927) was the first to develop the formal theory for equilibrium frequencies of alleles in mutation–selection balance, and these calculations later formed the basis for the indirect estimates of mutation rate in his 1935 paper. Haldane suggested that if selection acts against deleterious alleles it must be balanced by mutation pressure to generate observed allele frequencies in a population. This provided a straightforward way to estimate the mutation rate for deleterious alleles: if the frequency of an allele could be measured and if the strength of selection could be estimated, it should be possible to calculate the mutation rate. Since most mutations in genes are harmful, the mutation rate for deleterious alleles will be only a slight underestimate of the total mutation rate. For autosomal dominant mutations, the equilibrium allele frequency, q, is simply μ/s, where μ is the mutation rate and s is the selection coefficient associated with the deleterious mutant. This well-known result can now be found in any population genetics textbook (e.g. Crow and Kimura 1970, equation 6.2.8). The first application of this idea to estimate the mutation rate in humans was brief and came from Haldane in his 1932 book *The causes of evolution*, where he suggested that mutations causing haemophilia arise at a rate of roughly 10^{-5} per generation. Haemophilia is a recessive X-linked disorder, and in this case the mutation rate is given by $\mu = qs/3$, where q is the frequency of haemophiliac males in the population. If most haemophiliacs do not reproduce (i.e. if s is close to 1), then the mutation rate is roughly three times the frequency of male carriers (the denominator contains 3 since the gene is X-linked, and males carry one-third of the X chromosomes in the population). In his 1932 book, Haldane did not try to estimate the value of s, and thus he gave only an approximation for the mutation rate.

The first detailed study of the mutation rate for a gene under mutation–selection balance came three years later (Haldane 1935). This classic paper is sometimes cited as the first estimate of the mutation rate in humans, but this is not entirely accurate. As recognized by Muller (1950), Danforth (1923) had come up with a similar approach 12 years earlier to estimate mutation rates for syndactyly and polydactyly in humans. Like Haldane, Danforth understood that observed allele frequencies might represent a balance between mutation and selection. He reasoned that the frequency and persistence time of harmful alleles could therefore be used together to provide an estimate of the mutation rate. However, without accurate estimates of persistence time (which vary depending on how harmful a mutation is), Danforth was only able to provide an up-

*E-mail: nachman@u.arizona.edu.

Keywords. mutation; haemophilia; mutation–selection balance.

per bound for the mutation rate, which he estimated to be approximately 2×10^{-4} for syndactyly. Danforth understood that mutation rates were likely lower than this, but it remained unclear by how much his calculation was an overestimate of the true value.

Haldane's 1935 paper is a clear and comprehensive treatment of the problem. Using estimates of the frequency of haemophilic men in London, he was able to provide a good estimate of the frequency of the harmful allele. This approach works easily for X-linked recessives since they are visible in males. The paper is well written and direct. The main argument is nicely summarized in four sentences:

> If x be the proportion of haemophilic males in the population, and f their effective fertility, that is to say their chance, compared with a normal male, of producing offspring, then in a large population of $2N$, $(1–f)xN$ haemophilia genes are effectively wiped out per generation. The same number must be replaced by mutation. But as each of the N females has two X-chromosomes per cell, and each of the N males one, the mean mutation rate per X-chromosome per generation is $1/3(1–f)x$, or if f is small, a little less than $1/3(x)$. Hence we have only to determine the frequency of haemophilia in males to arrive at the approximate mutation rate. [p. 318]

In the commonly used current notation, $s = 1–f$. Haldane carefully reviewed the available evidence on fertility rates for haemophiliacs and concluded that f was probably between 0.1 and 0.25. He then reviewed the available data on the frequency of haemophilic males in London and concluded that x 'quite certainly exceeds 10^{-5}, and probably lies between 0.00004 and 0.00017' (p. 322). From this, he concluded that the mutation rate probably lies between 1 in 100,000 and 1 in 20,000, suggesting 2×10^{-5} as a plausible figure.

How does Haldane's estimate compare with subsequent research? The remarkable truth is that, years before anyone knew the structure or identity of genes, Haldane's estimate was very accurate. A good review of mutation rates estimated from many genes, including those underlying haemophilia, is given by Vogel and Motulsky (1997). For many genes, mutation rates are on the order of 10^{-5} per locus.

There are two main approaches for estimating mutation rates in humans, in addition to the method pioneered by Haldane. The first involves direct counts of affected individuals born to unaffected parents for autosomal or X-linked dominant disorders or for recessive X-linked disorders with severe effects. This method is simple and has been used for many diseases, and mutation rates vary over about two orders of magnitude from roughly 10^{-6} to 10^{-4} per locus per generation (Vogel and Motulsky 1997). By sequencing alleles in affected individuals, it is also possible to estimate mutation rates per nucleotide site (e.g. Sommer 1995; Kondrashov 2003). Methods that rely on screens of visible phenotypes almost certainly provide an underestimate of the true mutation rate at a gene, since some mutations will not produce a disease phenotype. This bias can be taken into account by estimating the fraction of sites in a gene at which mutations will produce the disease phenotype. This approach yields an overall estimate of mutation rate per nucleotide site of about 2×10^{-8} (Kondrashov 2003).

The second method is indirect and is based on the result that for neutral mutations the mutation rate is equal to the rate of evolution (i.e. the rate of fixation in a population per generation) (Kimura 1968). Thus, stretches of noncoding DNA can be compared between two species to calculate the amount of sequence divergence. If the generation time and the time since the species diverged are known, the mutation rate per generation can be estimated. This approach has been used for synonymous sites and pseudogenes in comparisons between human and chimpanzee, and these studies suggest mutation rates of about $1–2 \times 10^{-8}$ per nucleotide site (Kondrashov and Crow 1993; Drake *et al*. 1998; Nachman and Crowell 2000). Thus, the direct and indirect estimates are in very good agreement. Furthermore, since many genes have about 10^3 sites, these estimates correspond well with per-locus rates of 10^{-5}, as originally calculated by Haldane.

In subsequent work, Haldane (1947) suggested that more mutations may come from the male germ line than from the female germ line, a result that has since been well supported by molecular studies (Makova and Li 2002; Wolfe and Li 2003).

One of the more recent findings is the considerable variation in mutation rate at different classes of sites. For example, single-nucleotide substitutions are at least one order of magnitude more common than other types of mutations, such as insertions or deletions, and single-nucleotide substitutions at CpG sites are much more frequent than at other sites (Sommer 1995; Nachman and Crowell 2000; Kondrashov 2003). In some cases, the mechanistic basis for these differences is understood. For example, the high rate of transitions at CpG sites in mammalian genomes is the result of deamination of 5-methylcytosine to produce thymine (Cooper and Krawczak 1993). Recent work also suggests that mutation rates may differ in different regions of the genome (Wolfe and Li 2003). With the completion of the human genome sequence (International Human Genome Sequencing Consortium 2001), we can now estimate the mutation rate per genome more accurately, and it appears that each individual harbours roughly 100–200 new mutations. Since protein-coding genes constitute a small proportion of the human genome, most of these mutations will not occur in genes. However, one of the surprises of genomics has been the large numbers of noncoding sites that are conserved between species and therefore presumably have important functions (e.g. Kondrashov and Shabalina 2002; Mouse Genome Sequencing Consortium 2002; Boffelli *et al*. 2003;

Dermitzakis *et al*. 2003, 2004). Some nontrivial fraction of new mutations will fall into such regions. The mutations that occur in genes may therefore represent only a subset of the important mutations in humans, a situation that Haldane could not have foreseen.

Haldane's contribution to our understanding of the human mutation rate was important in at least two respects. First, he clearly articulated a method for estimating the mutation rate that could be used in humans. Second, he got it right: his estimate was remarkably accurate. Haldane of course made many fundamental contributions to genetics, and these are all the more impressive in light of the fact that he never received a scientific degree and had almost no formal training in biology (Provine 1971, p. 168). Among his many contributions to the field was the fact that he edited this journal for quite a few years; it is thus particularly appropriate that his highly influential 1935 paper be reproduced here.

Acknowledgements

I thank Matt Dean for comments on the manuscript.

References

Boffelli D., McAuliffe J., Ovcharenko D., Lewis K. D., Ovcharenko I., Pachter L. and Rubin E. M. 2003 Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299**, 1391–1394.

Cooper D. N. and Krawczak M. 1993 *Human gene mutation*. Bios Scientific, Oxford.

Crow J. F. and Kimura M. 1970 *An introduction to population genetics theory*. Burgess, Minneapolis.

Danforth C. H. 1923 The frequency of mutation and the incidence of hereditary traits in man. In *Eugenics, genetics and the family, Scientific papers of the 2nd International Congress of Eugenics, N.Y., 1921* **1**, 120–128. Williams & Williams, Baltimore.

Dermitzakis E. T., Reymond A., Scamuffa N., Ucla C., Kirkness E., Rossier C. and Antonarakis S. E. 2003 Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs). *Science* **302**, 1033–1035.

Dermitzakis E. T., Kirkness E., Schwarz S., Birney E., Reymond A. and Antonarakis S. E. 2004 Comparison of human chromosome 21 conserved nongenic sequences (CNGs) with the mouse and dog genomes shows that their selective constraint is independent of their genic environment. *Genome Res*. **14**, 852–859.

Drake J. W., Charlesworth B., Charlesworth D. and Crow J. F. 1998 Rates of spontaneous mutation. *Genetics* **148**, 1667–1686.

Haldane J. B. S. 1927 A mathematical theory of natural and artificial selection. Part V. Selection and mutation. *Proc. Cambridge Philos. Soc*. **23**, 838–844.

Haldane J. B. S. 1932 *The causes of evolution*. Longmans, Green & Co., London.

Haldane J. B. S. 1935 The rate of spontaneous mutation of a human gene. *J. Genet*. **31**, 317–326.

Haldane J. B. S. 1947 The mutation rate of the gene for haemophilia, and its segregation ratios in males and females. *Ann. Eugen*. **13**, 262–271.

International Human Genome Consortium 2001 Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.

Kimura M. 1968 Evolutionary rate at the molecular level. *Nature* **217**, 624–626.

Kondrashov A. S. 2003 Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Hum. Mutat*. **21**, 12–27.

Kondrashov A. S. and Crow J. F. 1993 A molecular approach to estimating the human deleterious mutation rate. *Hum. Mutat*. **2**, 229–234.

Kondrashov A. S. and Shabalina S. A. 2002 Classification of common conserved sequences in mammalian intergenic regions. *Hum. Mol. Genet*. **11**, 669–674.

Makova K. D. and Li W.-H. 2002 Strong male-driven evolution of DNA sequences in humans and apes. *Nature* **416**, 624–626.

Mouse Genome Sequencing Consortium 2002 Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562.

Muller H. J. 1928 The measurement of gene mutation rate in *Drosophila*, its high variability, and its dependence upon temperature. *Genetics* **13**, 279–357.

Muller H. J. 1950 Our load of mutations. *Am. J. Hum. Genet*. **2**, 111–176.

Nachman M. W. and Crowell S. L. 2000 Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**, 297–304.

Provine W. B. 1971 *The origins of theoretical population genetics*. University of Chicago Press, Chicago.

Sommer S. S. 1995 Recent human germ-line mutation: inferences from patients with hemophilia B. *Trends Genet*. **11**, 141–147.

Stadler L. J. 1932 On the genetic nature of induced mutations in plants. *Proceedings of the 6th International Congress of Genetics*, **1**, 274–294. Brooklyn Botanic Garden, Menasha, USA.

Vogel F. and Motulsky A. G. 1997 *Human genetics: problems and approaches*. Springer, Berlin.

Wolfe K. H. and Li W.-H. 2003 Molecular evolution meets the genomics revolution. *Nat. Genet*. **33**, 255–265.