# Mutational bias provides a model for the evolution of Huntington's disease and predicts a general increase in disease prevalence

David C. Rubinsztein[1], William Amos[2], Jayne Leggo[1], Sandy Goodburn[1], Rajkumar S. Ramesar[3], John Old[4], Ronald Bontrop[5], Robert McMahon[1], David E. Barton[1] & Malcolm A. Ferguson-Smith[1,6]

Huntington's disease (HD) correlates with abnormal expansion in a block of CAG repeats in the Huntington's disease gene. We have investigated HD evolution by typing CAG alleles in several human populations and in a variety of primates. We find that human alleles have expanded from a shorter ancestral state and exhibit unusual asymmetric length distributions. Computer simulations are used to show that the human state can be derived readily from a primate ancestor, without the need to invoke natural selection. The key element is a simple length-dependent mutational bias towards longer alleles. Our model can explain a number of empirical observations, and predicts an ever-increasing incidence of HD.

[1]East Anglian Regional Genetics Service Molecular Genetics Laboratory and Department of Clinical Genetics, Box 158, Addenbrooke's NHS Trust, Hills Road, Cambridge CB2 2QQ, UK
[2]University of Cambridge Department of Genetics, Cambridge CB2 3EH, UK
[3]Department of Human Genetics, University of Cape Town, Observatory, 7925, South Africa
[4]Institute of Molecular Medicine, John Radcliffe Hospital, Oxford OX3 9DU, UK
[5]Medical Biological Laboratories TNO, PO Box 5815HV, 2280, Rijswijk, The Netherlands and
[6]University of Cambridge Department of Pathology, Cambridge CB2 1QP, UK

Correspondence should be addressed to D.C.R.

Huntington's disease is one of a growing number of genetic disorders which are now known to be associated with expansion within a block of trinucleotide repeats[1]. Other examples include fragile X, Kennedy's disease and myotonic dystrophy[1]. The repeats responsible for HD have a CAG motif and are located in the 5' end of the transcript[1]. HD is a late onset disease, generally affecting people who are past reproductive age, and results when the CAG repeats exceed 34 in number[2]. When exceptionally large numbers of repeats are present, the disease manifests itself earlier in life[3–6].

HD is generally rare, but varies in prevalence between ethnic groups, being more common in East Anglians (UK) (1 in 10,000) and relatively rare amongst Japanese (1 in a million)[7]. As yet, no clinical symptoms have been linked to people carrying non-disease alleles (<35 repeats), suggesting that non-HD alleles do not confer overtly variable fitness. The lowest number of CAG repeats yet recorded in humans is eight, the individual(s) concerned being apparently healthy[2]. Why alleles shorter than this have not been observed remains to be established. The possibility that short alleles are selected against is difficult to rule out, although our observation that non-human primates all carry short alleles (see below) provides circumstantial evidence that this is not so.

Overall, therefore, it seems that natural selection acts weakly, if at all, on HD alleles below the disease threshold. In the absence of natural selection, the distribution of HD alleles in human populations should be explicable in terms of a balance between the mutational mechanisms generating new allele lengths and neutral genetic drift. In order to gain insight into the mutational mechanisms responsible, and to try to understand the evolution of disease chromosomes, we have analysed the distribution of non-HD repeat lengths in both human populations and primates. These results are compared with computer simulations modelling the mutational process. We conclude that human CAG repeats are not in equilibrium. Instead the CAG repeats appear to be subject to a mutational bias which causes inexorable expansion and will in turn lead to an ever-increasing incidence of Huntington's disease.

## Human CAG allele distributions

The block of CAG repeats which is responsible for HD lies immediately adjacent to a block of CCG repeats[6,8]. CAG repeat length has been determined by PCR, generally using primers which flank both the CAG and the CCG repeats[1]. However, since the CCG repeats are themselves polymorphic, we elected to determine the length of each block separately, using an internal PCR primer[6,8,9]. In this way, precise CAG repeat numbers were determined for samples drawn from five ethnic groups (Fig. 1, Sub-Saharan Blacks, East Anglians (our local population in the UK), Asians, Japanese and South African Blacks, see Methodology.

Two distinctive features are apparent. First, there is an apparent asymmetry to all distributions, more alleles lying above, rather than below, the modal length. All populations show positively skewed distributions (see legend, Fig. 1), similar to that seen at the *CEB1* VNTR locus[10]. At *CEB1*, the asymmetry is associated with mutational bias[10], leading us to wonder whether mutational bias could also explain the distribution of CAG alleles.

Second, although modal CAG repeat length is relatively consistent between human populations (either 15 or 17),

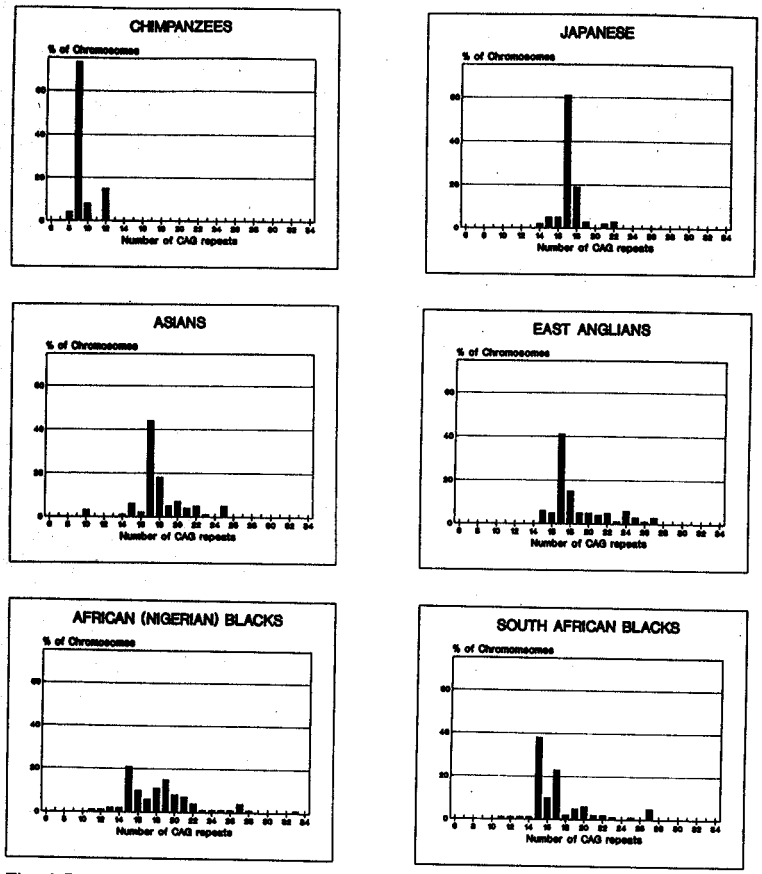


Fig. 1 Population distributions of CAG repeats in the Huntington's disease gene in five human populations and in chimpanzees. Skewness was determined, using the SPSS/PC+ package (see Methodology), to be 1.9 for the Africans (Nigerians), 4.0 for the East Anglians, 4.0 for the Asians, 4.5 for the Japanese, 3.4 for the South African Blacks and 2.0 for the chimpanzees. A skewness value of 0 indicates a normal distribution, while positive values indicate positive skews. The analysis of all of these populations and of the human populations alone revealed significant heterogeneity ($\chi^2$ = 424, 45 d.f., $p<0.0001$ and $\chi^2$ = 144, 36 d.f., $p<0.0001$).

the degree of spread varies, being greatest amongst the African groups and least amongst the Japanese. This variation is consistent with the population histories of the ethnic groups concerned. The Japanese sample shows low allelic diversity with a significant paucity of long alleles (>22 repeats, standard residual = –2.2, $p<0.05$). This is to be expected of a race which is thought to have gone through a relatively recent bottleneck, expanding from its founders around 300 B.C.[11]. Similarly, the African samples show much greater allelic diversity. This is compatible with expectations for a sample containing chromosomal lineages drawn from several ancient, previously isolated populations.

## Determination of the ancestral state

To determine the ancestral state, we extended our study to cover a wide range of primates (Fig. 1, Fig. 2). In total, sample individuals from ten species were typed, including 13 chimpanzees, the species thought by many to be man's closest living relative. CAG repeat lengths were remarkably consistent between species, with the combined range spanning only six repeat units (range 7–12). In size, the primate repeat numbers lie at the very bottom end of the human range, suggesting strongly that human CAG repeats have expanded from a shorter ancestral state.

## Linkage disequilibrium in CAG-CCG haplotypes

Use of an internal primer allows us to construct CAG-CCG haplotypes (Table 1, 2). CCG alleles are named relative to the published sequence, +0 being the same length, –1 being one repeat shorter, etc. Allelic diversity amongst CCG alleles is much lower than for CAG repeats, indicating a lower overall mutation rate. Interestingly, by typing other primates we found evidence that, like the CAG repeats, the CCG repeats have expanded during the course of human evolution. Most of the primates carry the +1 allele, and man's closest relatives, the chimpanzee and the gorilla, carry mainly –1 and +0/+1 alleles respectively.

As only 12 nucleotides separate these two regions[1], we expected and found strong linkage disequilibrium. For example, amongst East Anglians, the +0 allele is significantly over-represented in HD patients[6] (Table 1, $\chi^2$ = 13.07, 3 d.f., $p<0.005$). Amongst non-HD East Anglians, a significant linkage disequilibrium also exists between this CCG allele and long-normal CAG alleles (Mann Whitney U-Wilcoxon Rank Sum W test: U=46, W=24, $p<0.0458$). Indeed, all but one of 22 normal East Anglian chromosomes having 20 or more CAG repeats are associated with the +0 allele.

## Does meiotic drive affect HD alleles?

In myotonic dystrophy, another trinucleotide disease, heterozygous fathers carrying one allele with greater than 18 repeat units seem to preferentially transmit the longer allele to their offspring, leading to the suggestion that there may be meiotic drive at this locus[12]. To see if meiotic drive could also explain the increase in CAG repeat number in HD, we analysed a number of CEPH families involving parents who were heterozygous for CAG repeat number.

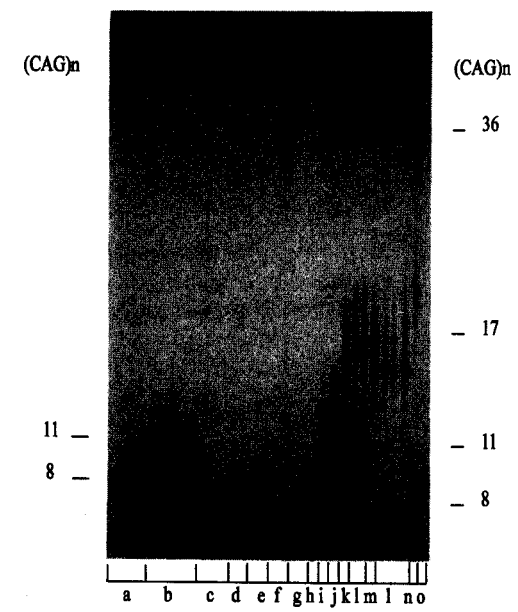


Fig. 2 CAG repeat sizes in the Huntington's disease genes of 3 gorillas (a), 3 baboons (b), 3 orang-utans (c), 2 crab-eating macaques (d), 2 rhesus macaques (e), 2 marmosets (f), pooled samples from 6 male talapoins or from 6 female talapoins (g), 1 gibbon (h) and 1 Olive baboon (i). Marker samples with 9 and 13 repeats (j) and 10 and 16 repeats (k), normal human samples (l), a marker indicating 36 repeats (n), an HD patient (m) and the water blank (o) are also included. The sizes of the repeats are shown.

**Table 1 CCG repeats in different populations**

| No . of CCG repeats (relative to published sequence) | Chimpanzees | Black South Africans | African (Nigerian) Blacks | East Anglians | Asians | Japanese | HD chromosomes |
|---|---|---|---|---|---|---|---|
| −1 | 25 (96) | 5 (6) | 1 (1) | | | | |
| +0 | | 19 (23) | 39 (41) | 46 (58) | 69 (68) | 40 (65) | 32 (91) |
| +1 | 1 (4) | 12 (14) | 1 (1) | 3 (4) | | | |
| +2 | | 5 (6) | 2 (2) | 6 (8) | 7 (7) | | 1 (3) |
| +3 | | 39 (46) | 53 (55) | 25 (31) | 26 (25) | 22 (35) | 2 (6) |
| +4 | | 3 (4) | | | | | |
| +6 | | 1 (1) | | | | | |

Sizes of CCG expansions in different populations relative to the originally published sequence that has seven uninterrupted CCG triplets[1]. Absolute numbers of chromosomes of each genotype are indicated and the percentage of each population's chromosomes with a given number of CCG repeats is shown in parenthesis. Analysis of all of these populations or of the normal human populations showed significant heterogeneity ($\chi^2$ = 499, 36 d.f., p<0.0001 and $\chi^2$ =109, 24 d.f., p<0.0001, respectively).

Over all meioses considered, no transmission bias in favour of longer alleles was found (maternal alleles transmitted, 83 long, 83 short; paternal alleles transmitted, 93 long, 101 short: statistics as in ref. 12, p=0.28). For a more direct comparison with the myotonic dystrophy result, we also looked for preferential inheritance of alleles with 17 or more repeats. Again, no significant deviation from one to one was found (maternal alleles, 17 short, 25 long, p=0.11; paternal alleles, 32 short, 24 long, p=0.14). Thus, meiotic drive does not seem to be a major factor governing the population genetics of HD. Furthermore, if meiotic drive were responsible for the gradual expansion of the CAG repeats, then one would expect to see allele distributions with negative skews, as opposed to the positive skews we have observed.

### Computer modelling of CAG repeat evolution

The evolution of human CAG repeats was modelled using a simple computer simulation. Each simulation begins with a single lineage in the assumed ancestral state (range tested, 9–13 repeats, chosen on the basis of the repeat sizes in non-human primates, see above) and continues for a

**Table 2 CAG/CCG haplotypes in different populations**

| CAG | Chimps (26 chromosomes) CCG | | | South African blacks (84 chromosomes) CCG | | | | | | | Africans (Nigeria) (96 chromosomes) CCG | | | | | East Anglians (80 chromosomes) CCG | | | | Asians (102 chromosomes) CCG | | | | Japanese (62 chromosomes) CCG | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | −1 | +0 | +1 | −1 | +0 | +1 | +2 | +3 | +4 | +6 | −1 | +0 | +1 | +2 | +3 | +0 | +1 | +2 | +3 | +0 | +1 | +2 | +3 | +0 | +1 | +2 | +3 |
| 7 | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 8 | 4 | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 9 | 73 | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 10 | 4 | | 4 | | | | | 1 | | | | | | | 1 | | | | | 1 | | | 2 | | | | |
| 11 | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 12 | 15 | | | 1 | | | | | | | | | | | 1 | | | | | | | | | | | | |
| 13 | | | | 1 | | | | | | | | 2 | | | | | | | | | | | | | | | |
| 14 | | | | | | | | 1 | | | | | | | 2 | | | | | 1 | | | | 2 | | | |
| 15 | | | | | | | 5 | 29 | 4 | 1 | | 1 | | 2 | 18 | 4 | | | 3 | 3 | | | 3 | | | | 5 |
| 16 | | | | 2 | 4 | | | 4 | | | 1 | 6 | | | 3 | 3 | | | 3 | 1 | | | 1 | 3 | | | 2 |
| 17 | | | | 1 | 11 | 7 | | 4 | | | | 4 | | | 2 | 19 | 4 | 8 | 11 | 29 | | 5 | 10 | 35 | | | 26 |
| 18 | | | | | 1 | | | 1 | | | | 4 | | | 7 | 4 | | | 11 | 11 | | 2 | 5 | 16 | | | 3 |
| 19 | | | | | 1 | 4 | | | | | | 8 | | | 6 | 3 | | | 3 | 3 | | | 2 | 3 | | | |
| 20 | | | | | 4 | | | 2 | | | | 4 | | | 4 | 5 | | | | 6 | | | 1 | | | | |
| 21 | | | | | | | | 2 | | | | 3 | | | 4 | 4 | | | | 4 | | | | 2 | | | |
| 22 | | | | | 1 | | | 1 | | | | 2 | | | 2 | 5 | | | | 5 | | | | 3 | | | |
| 23 | | | | | | | | 1 | | | | 1 | | | | 1 | | | | | | | 1 | | | | |
| 24 | | | | | | | | | | | | 1 | | | | 5 | | | 1 | | | | | | | | |
| 25 | | | | | 1 | | | | | | | 1 | | | | 3 | | | | 4 | | | 1 | | | | |
| 26 | | | | | | | | | | | | | 1 | | | 1 | | | | | | | | | | | |
| 27 | | | | | | 4 | 1 | | | | | 2 | | | 2 | 3 | | | | | | | | | | | |
| 28 | | | | | | | | | | | | | | | 1 | | | | | | | | | | | | |
| 29 | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 30 | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 31 | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 32 | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 33 | | | | | | | | | | | | | | | 1 | | | | | | | | | | | | |

CAG/CCG haplotypes in five different human populations and in chimpanzees. All figures represent the percentage of the chromosomes on the relevant haplotype. The CCG repeat-sizes are named relative to the number of repeats in the original published sequence[1]. The most common human CCG alleles (+0 and +3) are shaded for ease of identification.

total of 100 steps, 1 step being equivalent to a large number of actual generations. At each step, a lineage may either mutate (probability range tested, 0.01–0.15 in the ancestral state) or give rise to a sister lineage (probability range tested, 0.02–0.15; high values lead to more branching). Mutations all involve the gain or loss of a single repeat unit, despite the fact that larger bite sizes have been observed. This simplifying assumption was made in the expectation that larger bite sizes will merely increase allele size variance. The model is stochastic, all events are determined by selecting random numbers. For example, if the bias is 0.8, a random number between 0 and 1 is drawn and, if less than 0.8 a repeat unit is gained, if greater than 0.8 a repeat unit is lost. Simulations are stopped as soon as 1000 lineages have attained 100 steps. Where the first tree falls short of 1000 branches, further trees are initiated. All simulation conditions are repeated three times, mean allele frequencies noted and these means then compared by the least squares method for goodness of fit to the empirical distributions of all ethnic groups, and to the complete data set.

The hypothesis we are keen to test is that the human CAG allele distributions result from a mutational bias towards longer repeats. Such a bias was introduced to the model by varying the probability of gain rather than loss (probability range tested, 0.5–1.0 in the ancestral state). Both mutation rate and mutation bias were also given varying degrees of length dependency such that both reach maximal values in the disease range of allele lengths (exponential function varied between 1 and 6).

Using a wide range of starting conditions, we were able to generate a number of allele-length frequency distributions of a similar nature to those seen in humans. A representative distribution is presented in Fig. 3*a*, compared to the African sample (since this population consistently yielded a better fit than the other populations, see discussion below). In addition, example graphs are given to indicate the effect of individual parameters on the goodness of fit (Fig. 3*b*). Two patterns emerge: first, an essential feature of the model was found to be a mutational bias amongst all allele lengths. This is logical, since all human alleles apparently derive from a shorter, primate precursor. This would be very hard to explain without some force favouring longer alleles. Second, linear or weak logarithmic length dependency produces a poorer fit than strong length dependency. This might indicate a step function where the mutation rate and bias rise suddenly above a certain threshold length.

The main difference between the results of our simulations and the empirical distributions is that the allele frequencies in human populations are much spikier, certain alleles being over-represented. This is to be expected. Population sub-division, random genetic drift and migration will all affect the pattern we observe today. Particularly during the period when human populations were measured in thousands or tens of thousands, bottlenecks and genetic drift would allow some alleles to reach high frequency.

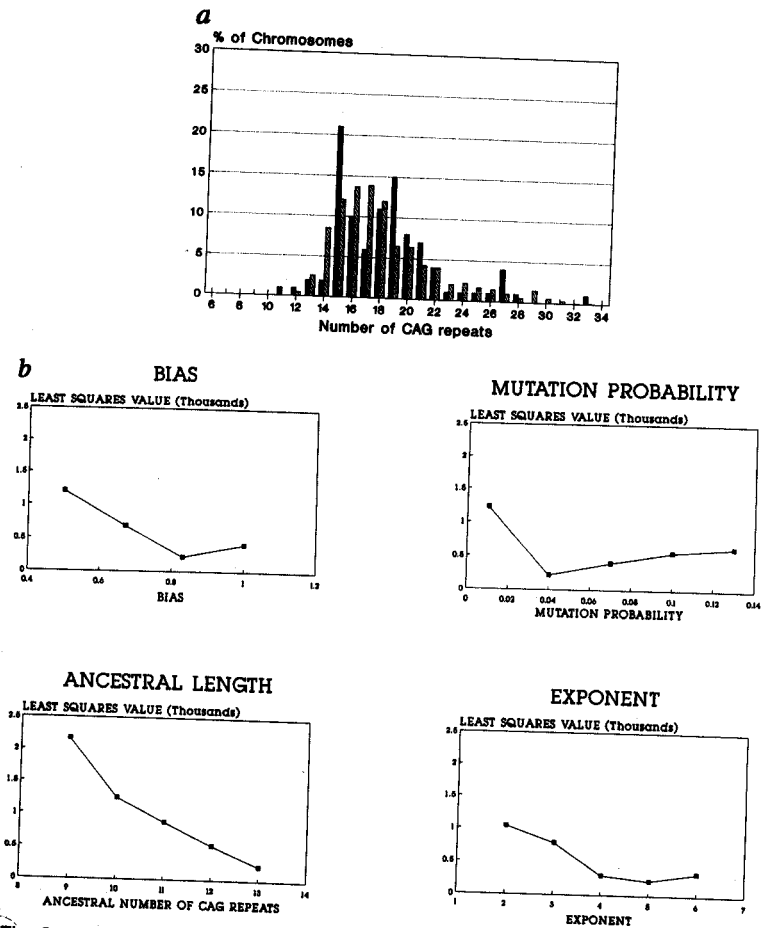


Fig. 3 *a*, Comparison between human CAG allele frequencies and allele frequencies derived from a computer simulation. 500 simulations were performed. A consensus set of best-fit parameters was derived ($\mu_s$ = 0.04, bias = 0.83, ancestral length = 13, length dependency exponent = 5). These settings produced the best fit in each population. However, the African (Nigerian) sample gave better fits to the simulated data for all settings. Therefore, we have modelled our experiment in Fig. 3*b* relative to the African (Nigerian) population. A comparison of the African (Nigerian) data (filled bars) to the computer generated allele frequencies using the best-fit parameters (hatched bars) is shown. *b*, Influence of parameter settings on the goodness of fit of the computer simulation to the African (Nigerian) allele frequencies. In each case, three parameters are held constant at their best-fit values and the fourth is varied over the range tested.

## Discussion

**Mutational bias: a model for repeat expansion.** Huntington's disease is associated with abnormal expansion within a block of CAG repeats in the Huntington's disease gene[1]. Such selection explains the rarity of HD alleles containing large numbers of CAG repeats. Here we examine the possible forces acting to shape the distribution of non-HD allele lengths.

We have determined CAG repeat lengths for sample alleles drawn from five different human populations, as well as from a wide variety of primate species. The allele lengths we find show a number of interesting and unusual properties which must be explained by any model addressing evolution of HD. First, virtually all human chromosome lineages carry more CAG repeats than are likely to have been present in the ancestral state. Second, the human allele-frequency distributions all reveal a distinct skew towards longer alleles. Third, whilst the modal allele length varies little, variance in allele length often differs between populations. Finally, disease incidence varies greatly between ethnic groups[7].

The above observations can be explained by invoking natural selection, either favouring the current mode or eliminating short alleles. Whilst there is no evidence to support this possibility, it is difficult to preclude. We have

therefore decided to follow an alternative approach. If selection is assumed to be absent, a simple model based on mutation and neutral genetic drift should be capable of generating the allele frequency distributions we see today. The key feature of such a model, mutational bias, is suggested by three observations: the general expansion in CAG repeat number, the bias which is known to operate on disease length alleles and the distribution of allele lengths at a minisatellite locus where mutational bias has been documented[10].

**Computer simulations of CAG repeat evolution.** We used computer simulations, incorporating both mutational bias and length dependency, to investigate whether the observed distribution of HD alleles in different ethnic populations can be explained by drift and mutation alone. Our simulations proved able to generate allele frequency distributions of a very similar form to the empirical data. The main discrepancy is that the simulated distributions are smoother and fail to match the low variance of the Japanese sample.

This shortcoming reflects a deficiency in the model, and is to be expected. We cannot hope to model past, unknown stochastic events in human evolution. Instead, our model produces a probability distribution for chromosomal end-states in a very large population. Smaller populations, and those which have experienced a bottleneck, will show lower allelic diversity and, during expansion, will have a lower variance for allele length. In other words, the model is good at predicting which alleles are likely to be commonest, but it does not incorporate the sort of founder effects which would allow individual alleles to reach high frequency.

This argument predicts that our model should fit best to a mixture of many ancient lineages, and should fit worst to a pure population which has passed through a recent bottleneck. This is exactly what we find. Best fit to the simulation is achieved by populations which could be argued to contain diverse lineages, such as the Africans and the pooled data set, and poorest fit is achieved by the Japanese, who are thought to have experienced a recent bottleneck.

The tendency of our model to inflate variance in allele size relative to natural populations has an important consequence. In the best fit solution, parameter settings which reduce allele length variance will be set artificially high. For example, allele length variance is inversely proportional to the degree of mutational bias. Consequently, although mutational bias is an absolute requirement, the degree of bias suggested by the model is probably too high. Similarly for ancestral length, lineages which are initiated close to the empirical mode will require fewer mutational steps, show a lower variance and tend to fit the empirical data better. We thus wish to emphasise that our model is not designed to predict precise values for each of the parameters tested, but more to demonstrate that the human state can be derived from its presumed ancestral state.

As a further example of the qualitative nature of the model, consider mutational bias. We assume that all mutations involve a single repeat unit, bias operating simply to determine whether this is gained or lost. Our observations do not preclude very rare jumps of many repeat units. The positive skewing of the distributions that we have noted is largely created by alleles close to the modal length and is unlikely to be due to very large jumps. However, bite sizes of more than one repeat have been observed[13]. An alternative possibility is that bias correlates with bite size, such that single repeat units are gained and lost with equal frequency but larger blocks are more likely to be gained rather than lost. This possibility is entirely speculative, but would provide a possible means to link bias, mutation rate and length dependency.

A neutral model gains further support from the strong linkage disequilibrium which exists between neighbouring CAG and CCG repeats. Compared with a presumed ancestral state amongst primates, both these repeats appear to be expanding, but at different rates, the CCG repeats being slower. Low CCG repeat number alleles should therefore indicate ancient chromosomal lineages and higher CCG repeat numbers will mark more recent lines. Since the ancient lineages will have had a longer time for the CAG repeats to expand, we would predict that low CCG repeats should be associated with higher mean CAG repeat lengths and *vice versa*. This is precisely what is found, CCG and CAG repeat lengths show a highly significant negative correlation ($r=0.23$, 448 d.f., $p<0.001$). Our data is also compatible with the alternative possibility that a jump from CCG +0 to CCG +3 occurred in human evolution to give rise to the bimodal distribution of this repeat. This scenario would also give rise to the disequilibrium noted above.

Finally, a further prediction of our model is that the prevalence of HD in a population should correlate with the frequency of long-normal alleles (20–33 repeats). Recent work has shown that new HD mutations originate on chromosomes with high normal repeat numbers[14,15]. HD varies greatly in prevalence between races. Amongst East Anglians, HD is relatively common[7]. As predicted, East Anglians also carry many more long-normal alleles, in this case associated with a putative ancient +0 CCG allele. That this is an effect of linkage disequilibrium and genetic drift, and is not merely an effect of the +0 allele itself, is shown clearly by the fact that Japanese people carry many +0 CCG alleles, yet these are not particularly associated with long-normal CAG alleles. Similarly, HD is very rare amongst Japanese, and this race has a significant deficit of long-normal alleles relative to all other races.

In conclusion, we have constructed a simple model which can explain readily the distribution of trinucleotide repeats in the human Huntington's disease gene. The model indicates two requirements, a mutational bias towards longer alleles over all allele lengths and a strong length dependency for the mutation rate, possibly even a step-function. Although the precise parameters must be determined empirically, our model can explain many unusual features of the HD CAG repeats: their evolution from an ancestral primate state, the asymmetric distribution of allele lengths, the different prevalences of the disease in different populations and the nature of the linkage disequilibrium between the CAG and CCG repeats. The model is simple, and consistent with other observations on the mutational process in the Huntington's disease gene.

If our model holds true, there is an unfortunate consequence that has recently also been considered by John Maddox[16]. We predict that the allele frequency distributions we see today are not end-points, but a snap-shot of a transition state in an on-going process of gradual expansion. In the absence of interference, this expansion

will continue and accelerate, leading to an ever-increasing incidence of HD. Since mutation towards detrimental repeat lengths happens over the course of several generations, natural selection will be entirely ineffective in halting this process.

*Note added in proof:* We have recently found two chimpanzee chromosomes with 16 CAG repeats. CAG repeat lengths were determined in a further 12 chimpanzees (three known founders and nine other chimpanzees of unknown origin — we cannot rule out relatedness of these animals either to each other or to the founders). Seventeen chromosomes had 9 repeats, one had 10 repeats, three had 12 repeats, one had 13 repeats and two had 16 repeats. The high prevalence of small repeats confirms our previous findings and conclusions that the ancestral state was a small number of repeats. The finding of chromosomes with 16 repeats suggests either that the expansion from the ancestral state occurred before the split of the human and chimpanzee lineages or that expansions have occurred in both lineages.

## Methodology

**DNA samples.** DNA was examined from 40 East Anglian cystic fibrosis patients who were not homozygous for the ΔF508 mutation, 48 Sub-Saharan Africans (mainly from Nigeria) who were referred for the diagnosis of sickle-cell anaemia, 51 Asian patients from the Indian subcontinent and Pakistan who were referred for diagnosis for haemoglobinopathies, 31 Japanese individuals who attended language schools in Cambridge, 42 normal South African Blacks (mainly Xhosa) and chimpanzees (*Pan troglodytes* from Sierra Leone). The East Anglians, Sub-Saharan Africans, Asians and South African Blacks did not deviate significantly from Hardy-Weinberg equilibrium when the observed homozygote and heterozygote frequencies of the insertion/deletion polymorphism in the myotonic dystrophy gene[17] were compared to expected values deduced from the observed allele frequencies ($\chi^2 < 0.05$ in all cases). Similarly, the Japanese sample did not deviate from Hardy-Weinberg equilibrium when tested with the genotypes observed for the the HD CCG repeats ($\chi^2 < 0.05$). The chimpanzees represent 13 founder individuals caught in the wild. This sample is relatively outbred as determined by their MHC haplotype heterogeneity[18]. In addition, 3 gorillas (*Gorilla gorilla*), 3 baboons (*Papio* sp.), 3 orang-utans (*Pongo pygmaeus*), 2 crab-eating macaques (*Macaca fascicularis*), 2 rhesus macaques (*Macaca mullato*), 2 marmosets (*Callithrix* sp.), 1 Olive baboon (*Papio anubis*), one gibbon (*Hylobates* sp.) and pooled samples either from 6 male or from 6 female talapoins (*Miopithecus* sp.) were analysed. All DNA samples were from peripheral blood lymphocytes except for the Japanese samples which were from mouthwashes and the orang-utan, gibbon and gorilla samples which were derived from cell lines.

Received 9 March; accepted 16 May 1994.

**PCR analysis.** The CAG and CCG repeats in the Huntington's disease gene were amplified either with primers that span each individual tract or with primers that span both repeats. CAG repeats were amplified as described[9] except that α-$^{32}$P-dCTP was incorporated and a 3:1 ratio of 7-deaza-dGTP to dGTP was used. Allele sizes were determined by comparison to a known sequence ladder. CCG repeats were amplified using one or both of the published methods[6,8] except that an annealing temperature of 66 °C was used for the non-human primates in the method of ref. 6. A further PCR was needed to determine the phase of each CAG repeat for each CCG repeat on every chromosome that has been reported here. As these two repeated sequences are adjacent to each other, we amplified across both repeats using the two flanking primers HD1 (5'–ATG AAG GCC TTC GAG TCC CTC AAG TCC TTC–3')[2] and Hu3 (5'–GGC GGC TGA GGA AGC TGA GGA–3')[19] and compared the result of this reaction to those for the CAG and CCG repeats alone. PCR reactions were performed in 25 μl containing 200 ng of each primer, 2 ng of 5' end-labelled Hu3, 1× PCR buffer (50 μM KCl, 10 mMTris (pH 8), 0.01% gelatin, 1.5mM $MgCl_2$) 10% DMSO, 200 μM dATP, dCTP and dTTP, 50 μM dGTP, 150 μM 7-deaza dGTP and about 1 μg DNA. Reactions (without the *Taq* polymerase) were heated for 5 min at 95 °C, 2.5 U of the polymerase (BRL) were added and samples transfered to a block that alternated between 95 °C for 1 min and 66 °C for 2 min for 35 cycles. In the case of the HD chromosomes the same protocol was used, except that the PCR described initially by the HD consortium[1] was used to deduce the CAG-CCG haplotype as it also includes both polymorphisms[6,8].

**Statistical analysis.** Statistics were performed using the SPSS/PC+ package (SSPC Inc., Chicago, Illinois). Distributions of the population frequencies of CAG and CCG repeats were examined using the $\chi^2$ test of independence. In order to examine where the differences lay within the populations, the residuals (i.e. the discrepancy between the observed and expected values) were analysed. As the sample size was large, the standardized residuals have an approximately normal distribution. A value of >1.96 or <–1.96 was taken as a 2 tailed level of p=0.05.

**Computer modelling.** The evolution of CAG repeats in the human HD gene was modelled using a computer simulation programmed in Turbo Basic. Details of the rationale for the model can be found in the Results section. An exponential length-dependent mutation probability was determined with the following formula $\mu = \mu_a + (1-\mu_a)L^x/33^x$, where $\mu$ is the mutation probability at a given CAG length (L), $\mu_a$ is the mutation probability in the ancestral state and $x$ is an exponent that was varied from 1–6.

**Acknowledgements**

*We thank Bronwen Lambson and Mike Hobart for providing DNA samples and P. Altham for statistical advice. R. Harding is thanked for her interesting discussion. This work was supported by the Huntington's Disease Association (UK) and the Royal Society (B.A.).*

1. The Huntington's Disease Collaborative Research Group. A Novel Gene Containing a Trinucleotide Repeat that Is Expanded and Unstable on Huntington's Disease Chromosomes. *Cell* **72**, 971–983 (1993).
2. Barron, L.H. *et al.* A study of the Huntington's disease associated trinucleotide repeat in the Scottish population. *J. med. Genet.* **30**, 1003–1007 (1993).
3. Duyao, M. *et al.* Trinucleotide repeat length instability and age of onset in Huntington's disease. *Nature Genet.* **4**, 387–392 (1993).
4. Andrew, S.E. *et al.* The relationship between trinucleotide (CAG) repeat length and clinical features of Huntington's disease. *Nature Genet.* **4**, 398–403 (1993).
5. Snell, R.G. *et al.* Relationship between trinucleotide repeat expansions and phenotypic variation in Huntington's disease. *Nature Genet.* **4**, 393–397 (1993).
6. Rubinsztein, D.C., Barton, D.E., Davison, B.C.C . & Ferguson-Smith, M.A. Analysis of the huntingtin gene reveals a trinucleotide-length polymorphism in the region of the gene that contains two CCG-rich stretches and a correlation between decreased age of onset of Huntington's disease and CAG repeat number. *Hum. molec. Genet.* **2**, 1713–1715 (1993).
7. Harper, P.S. The Epidemiology of Huntington's Disease. in *Huntington's Disease* (ed Harper, P.S.) 251–280 (Saunders Ltd, London, 1991).
8. Rubinsztein, D.C., Leggo, J., Barton, D.E. & Ferguson-Smith, M.A. Site of (CCG) expansion in huntingtin gene. *Nature Genet.* **5**, 214–215 (1993).
9. Warner, J.P., Barron, L.B. & Brock, D.J.P. A new Polymerase Chain Reaction (PCR) assay for the trinucleotide repeat that is unstable and expanded in Huntington's Disease chromosomes. *Molec. Cell Probes* **7**, 235–239 (1993).
10. Vergnaud, G. *et al.* The use of synthetic tandem repeats to isolate new VNTR loci: cloning of a human hypermutable sequence. *Genomics* **1**, 135–144 (1991).
11. Rouse, I. in *Migrations in Prehistory* 67–105 (Yale University Press, New Haven, London, 1986).
12. Carey, N. *et al.* Meiotic drive at the myotonic dystrophy locus? *Nature Genet.* **6**, 117–118 (1994).
13. Zuhlke, C., Reiss, O., Bockel, B., Lange, H. & Thies, U. Mitotic stability and meiotic variability of the $(CAG)_n$ repeat in the Huntington disease gene. *Hum. molec. Genet.* **2**, 2063–2067 (1994).
14. Goldberg, Y.P. *et al.* Molecular analysis of new mutations for Huntington's disease: intermediate alleles and sex of origin effects. *Nature Genet.* **5**, 174–179 (1993).
15. Myers R.H. *et al. De novo* expansion of a (CAG)n repeat in sporadic Huntington's disease. *Nature Genet.* **5**, 168–173 (1993).
16. Maddox, J. Triplet repeat genes raise questions. *Nature* **368**,685 (1994).
17. Mahadevan, M.S., Foitzik M.A., Surh, L.C. & Korneluk, R.G. Characterisation and polymerase chain reaction (PCR) detection of an *Alu* deletion polymorphism in total linkage disequilibrium with myotonic dystrophy. *Genomics* **15**, 446–448 (1993).
18. Slierendregt, B.L. *et al.* Major histocompatibilitycomplex class II haplotypes in a breeding colony of chimpanzees (*Pan troglodytes*). *Tissue Antigens* **42**, 55–61 (1993).
19. Reiss, O., Noerremoelle, A., Soerensen, S.A. & Epplen, J.T. Improved PCR conditions for the stretch of (CAG)n repeats causing Huntington's disease. *Hum. molec. Genet.* **2**, 637 (1993).