Analysis of the 5' upstream sequence of the Huntington's disease (HD) gene shows six new rare alleles which are unrelated to the age at onset of HD

Rhian Coles, Jayne Leggo, David C Rubinsztein

Abstract

The CAG repeat number in the Huntington's disease (HD) gene accounts for about 50% of the variation seen in age at onset of HD. In order to determine whether promoter sequence variation can contribute to the residual variation in age at onset, we studied the conserved 303 bp region upstream of the +1 translation start site in the HD gene in a population of 56 control East Anglians, 30 Africans, 34 Japanese, and 208 English Huntington's disease patients. A surprisingly high degree of variation was found. Seven alleles were identified, comprising four polymorphisms: two single base pair substitutions, a 6 bp VNTR present as one or two copies, and a 20 bp VNTR with one to three copies of the tandem repeat. No correlation between polymorphisms and age at onset of symptoms was found in HD patients. The 6 bp and 20 bp stretches are present only in single copies in the chimpanzees and gorilla, suggesting that these VNTRs have evolved by duplication of the core sequences in the human lineage.

(*J Med Genet* 1997;34:371–374)

Keywords: Huntington's disease; promoter; polymorphisms

Department of Pathology, Cambridge University, Cambridge, UK
R Coles

Department of Medical Genetics, Box 158, Addenbrooke's NHS Trust, Hills Road, Cambridge CB2 2QQ, UK
J Leggo
D C Rubinsztein

Correspondence to:
Dr Rubinsztein.

Received 11 October 1996
Revised version accepted for publication 19 December 1996

The Huntington's disease (HD) mutation is an expansion of a CAG repeat tract coding for a polyglutamine stretch near the N terminus of the huntingtin protein.[1] Normal subjects have 8-35 repeats, and expansion beyond 35 repeats is associated with the disease.[2] CAG repeat expansion is thought to result in the mutant protein acquiring some detrimental dominant gain of function.[3] A rough inverse correlation exists between the number of CAG triplets in the disease size range and the age at onset of symptoms,[1 4] with juvenile cases tending to have larger expansions than those with adult onset.[5] Repeat length appears to account for roughly 50% of the variation in age at onset and the HD mutation is occasionally not fully penetrant.[2] As a result, it has been suggested that genetic factors independent of the CAG repeat number also play a role in determining the age of onset of HD.[6]

HD is one of a group of diseases caused by polyglutamine expansion, including spinocerebellar ataxia type I (SCAI), dentatorubralpallidoluysian atrophy (DRPLA), spinobulbar muscular atrophy (SBMA), and Machado-Joseph disease (MJD). Transgenic mice expressing the mutant human SCA1 gene develop ataxia, and transgene mRNA levels in these lines were found to be at least 10-fold higher than endogenous murine *Sca-1*. Since two lines of *SCA1* transgenic mice had ataxia when bred to homozygosity, but remained asymptomatic in the heterozygous state,[7] levels of expression of the mutant allele may be a factor contributing to induction of neurodegeneration. Previous studies have shown no gross differences in levels of HD mRNA expression between patients and controls[8] and both normal and mutant alleles are expressed at roughly comparable levels in the cortex and striatum of two HD heterozygotes.[9] However, the possibility of polymorphisms in the HD promoter, which may subtly affect transcription, has not been considered.

We hypothesised that variation in the promoter region could affect levels of HD mRNA expression and this in turn could contribute to variation in the age at onset of disease symptoms, and so we screened part of the promoter for polymorphisms. A previous study of the promoter regions of the human and the mouse Huntington's genes showed conservation between the -56 and -206 positions (78.8% nucleotide identity), but this fell to only 50% further upstream.[10] Computer analyses identified consensus sequences for putative Sp1 and AP2 binding sites within this fragment which are conserved between man and mouse.[10] We therefore chose to study a 303 bp region extending upstream from the +1 translation start site, encompassing this highly conserved region, since this could potentially contain variants of functional importance.

Methods

DETECTION AND ANALYSIS OF POLYMORPHISMS

We used the following primers (with short overhanging tags): 5'-GCGCGAGCTCAG CGGCTTGCTGTGTGAGG-3' (forward, -323 to -304) and 5'-GCGCCTCGAG CTTTTCCAGGGTCGCCAT-3' (reverse, +1 to + 20) to PCR amplify the -1 to -303 bp region of the HD gene in a population of 56 apparently unrelated East Anglians, 30 Black Africans (mainly Nigerians), 34 Japanese, 208 English Huntington's disease patients for whom the age at onset was known, 28 chimps,

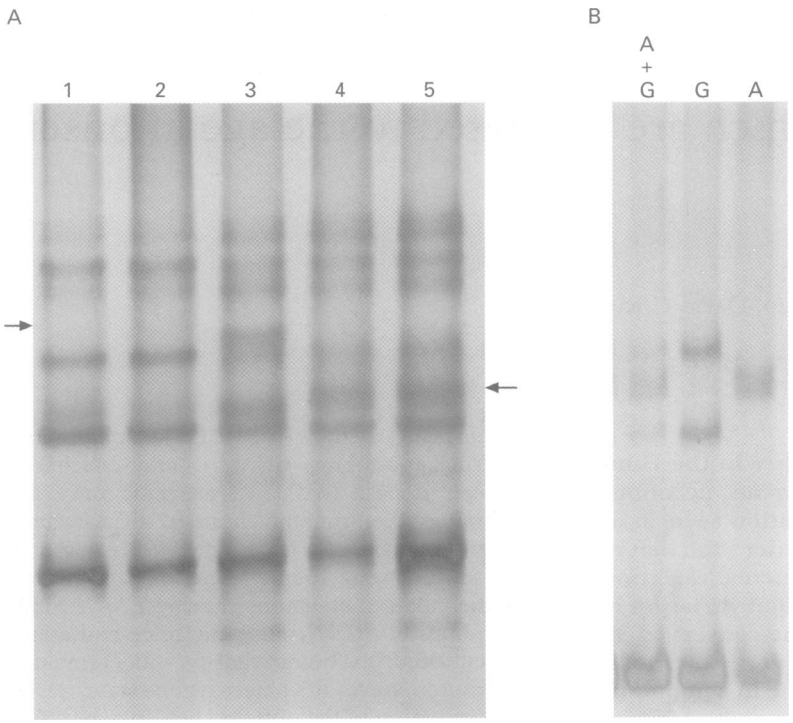

Figure 1 (A) SSCP analysis of a representative panel of HD patients. Arrows indicate additional bands produced by single base pair substitutions. Lane 3 has a C-T substitution at site IV (see fig 2) and lanes 4 and 5 have a G-A substitution at site III. Lanes 1 and 2 are wild type. (B) Cross species heteroduplex analysis of chimp DNA. Representative samples of the three types of intraspecific polymorphism are shown, with the bases present at site -173 indicated.

and one gorilla. Thermal cycling conditions were 95°C for two minutes, then 95°C for one minute 30 seconds, 64°C for one minute, 72°C for two minutes for 30 cycles. PCR products were subjected to SSCP analysis, by mixing 3 µl each of PCR product and formamide dye and denaturing at 95°C before placing immediately on ice. In addition, heteroduplex analyses were carried out, either by mixing 3.5 µl product with 3.5 µl non-denaturing dye, or by mixing 2 µl product with 6 µl PCR product of the commonest sequence and 8 µl non-denaturing dye, denaturing at 95°C for five minutes, then incubating at 70°C for one hour and cooling slowly to room temperature for one hour, to allow strands to reanneal. Samples for SSCP and heteroduplex analysis were loaded on 8% non-denaturing polyacrylamide gels (19:1 acrylamide:bisacrylamide) containing 5% glycerol, and run overnight at room temperature at 30 V. Gels were silver stained. A combination of these approaches showed a variety of variant banding patterns in the human populations, indicating that this fragment of the HD promoter is not monomorphic (fig 1A). Representative samples of each novel SSCP or heteroduplex pattern, numbering 48 in total, were first purified using the Wizard minipreps PCR purification system (Promega), then sequenced in both directions, manually using the Promega fmol kit or using the ABI Prism 377 automated sequencer, or both. For manual sequencing, primers were labelled with $\gamma^{32}P$; for automated sequencing, the Dye terminator cycle sequencing ready reaction kit (Perkin Elmer) was used.

STATISTICAL ANALYSIS

The possible genotypic effect of allele iii was determined by first studying the linear dependence of age at onset of HD on CAG repeat number. The best fit for our data was obtained by using the log transformation: log (age) = α + β (CAG repeat number) (data not shown). Residuals from this model were checked and there was no evidence of departure from normality and equality of variance assumptions (results not shown). The possible genotypic effect of allele iii was assessed using multiple linear regression, while allowing for the predictive effects of the CAG repeat size. Statistical analysis was performed using S-Plus.

Results

SSCP and heteroduplex analysis showed seven alleles within the -1 to -303 region of the human HD gene, relative to the translation start site (figs 1A and 2, table 1). These alleles resulted from variation at one or more of four polymorphic sites, labelled I-IV in fig 2. Two of these are single base pair substitutions: a G-A substitution at site III (allele ii) and a C-T at site IV (allele iii). The remaining two polymorphisms involve one or two copies of a 6 bp sequence, producing a perfect tandem repeat (site I) and one, two, or three copies of a 20 bp sequence (site II). The published, common allele (i) contains a single 6 bp tract at site I, a direct repeat of 20 bp at site II, a G at site III, and a C at site IV. Frequencies of each allele for the four populations are given in table 1.

Subjects who deviated from the originally published HD sequence at more than one site were rare, but included two Africans who were each homozygous for the 6 bp insertion and heterozygous for the G-A substitution at site III (having alleles iv and v), and one African who was homozygous for both of these rare variant sequences. One HD patient was heterozygous

Table 1 Characterisation of human HD promoter alleles and their frequencies in four populations

	Polymorphic site				*Frequency in population (No of alleles/total)*			
Allele	*I*	*II*	*III*	*IV*	*Africans*	*Japanese*	*East Anglians*	*HD patients*
i	1	2	G	C	0.683 (41/60)	0.897 (61/68)	0.902 (101/112)	0.904 (376/416)
ii	1	2	A	C	0.038 (2/52)	0 (0/42)	0.031 (3/96)	0.015 (6/400)
iii	1	2	G	T	0.038 (2/52)	0 (0/42)	0.021 (2/96)	0.0525 (21/400)
iv	2	2	G	C	0.167 (10/60)	0 (0/68)	0.018 (2/112)	0.0096 (5/416)
v	2	2	A	C	0.067 (4/60)	0 (0/68)	0 (0/112)	0.0024 (1/416)
vi	1	1	G	C	0 (0/60)	0 (0/68)	0.036 (4/112)	0.012 (5/416)
vii	1	3	G	C	0.017 (1/60)	0.103 (7/68)	0 (0/112)	0.0048 (2/416)

The number of 6 bp and 20 bp tandem repeats at sites I and II respectively are indicated, as are the bases at sites III and IV. All PCR products were subjected to heteroduplex analysis, and alleles i, iv-vii are expressed as a frequency of the total number of samples tested. Single base pair substitutions (alleles ii and iii) were only detected by SSCP (carried out on most samples) and are expressed as a frequency of the total analysed in this way.

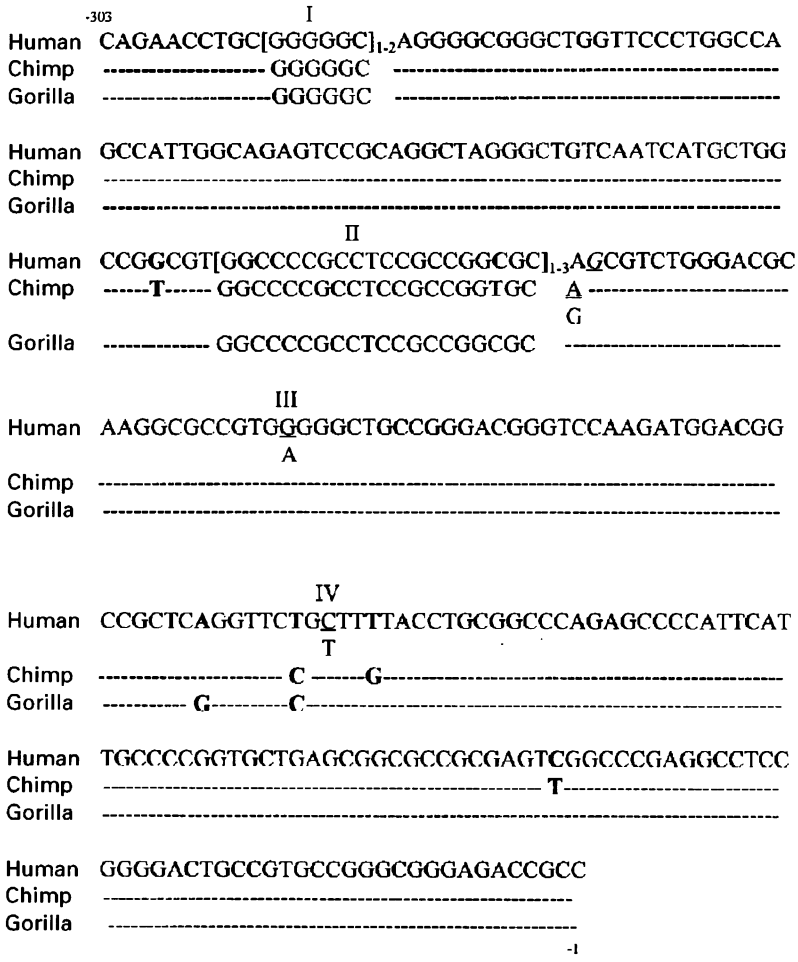

Figure 2 Sequence of the 5' upstream region of the HD gene in human, chimp, and gorilla. I-IV: sites found to be polymorphic in human populations. Dashes indicate identity of primate sequences with the human; bases which differ are highlighted in bold. The extra base found in all samples is underlined. Numbering is according to the +1 translation start site. These sequences have been submitted to EMBL. Accession numbers are: Y07981-7 (human variants), Y07988 (gorilla), and Y07989-90 (chimpanzee variants).

for the 6 bp insertion and homozygous for the G-A at site III (alleles ii and v). Another HD patient was heterozygous for both single base pair substitutions but the phase of these alleles could not be determined, since no family members were available.

Allele iii was the only variant common enough to warrant testing as a factor influencing age at onset of HD. However, the presence of this allele did not have a significant effect on variance in the age of onset unaccounted for by the CAG repeat number ($p>0.05$).

This region of the HD gene was highly conserved in man, chimp, and gorilla (fig 2). We observed 98.02% identity (six variants/303 bp, where the 20 bp tract is counted as a single variant) between the chimp and the commonest published human allele, while human and gorilla were 99% identical (three variants/303 bp) and gorilla and chimp 98.3% identical (five variants/303 bp). Both of the VNTRs found in humans (sites I and II) were present only as a single copy in all chimps and the gorilla. All samples, including humans with the commonest allele, chimps, and gorilla, had an extra G at position -172, compared with the published sequence.[10]

Initial SSCP and heteroduplex analyses of 28 chimps showed no apparent intraspecific variation (data not shown). However, when human PCR product with the published sequence was mixed with each chimp PCR product in a 3:1 ratio, different heteroduplex banding patterns were seen (fig 1B). Sequencing showed that the panel of chimps deviated at a single site (-173), and fell into three groups according to whether they were homozygous for an A, homozygous for a G, or heterozygous at this position. The cross species mixing could distinguish between these groups whereas heteroduplex confined to the individual chimp samples could not. We suggest that cross species heteroduplex analysis may be a more sensitive method for mutation detection in general.

Discussion

There is a comparative paucity of information regarding naturally occurring polymorphisms in the 5' upstream regions of human genes, with available data indicating low nucleotide diversity at these sites. No polymorphisms were found in 103 subjects for 500 bp upstream of the transcription start site in human complement component C4,[11] and a comprehensive comparison of randomly selected pairs of samples from 49 loci showed only a single nucleotide difference in a total of 3624 bp of 5' untranslated sequence.[12] Since bias towards reporting only highly variable sequences is probable, even this minimal extent of diversity may be over-represented. However, exceptions do occur, notably the highly polymorphic HLA class I and II alleles,[13–16] four variant sites occurring in the -220 to +1 region of the human cystatin C gene,[17] and nine polymorphisms in 814 bp of 5' flanking region of the β globin locus.[18] We believe that the four polymorphic sites (seven alleles) in the 303 bp upstream of the initiation codon of the Huntington's gene reflect a surprisingly high degree of variation. In addition, one of these polymorphic sites is thought to lie within the 5' untranslated region (site IV),[10] an apparently rare occurrence.[12 15] Despite the fact that both SSCP and heteroduplex analyses were used to screen for variation, it is possible that additional polymorphic sites may have remained undetected.

Of the reported polymorphisms, those involving single base pair substitutions are by far the most common.[16–18] The substitutions in this region of the HD promoter do not create or delete any consensus sequences for putative DNA binding proteins. However, two of the polymorphic sites identified in this study involved variable numbers of perfect repeats of six and 20 bp. Each 20 bp repeat contains a number of potential *cis* acting sequences, including an Sp1, AP2, and IRE motif. It is possible that there would be increased binding of *trans* acting factors when this stretch is duplicated or triplicated. No binding sites were identified within or around a single 6 bp stretch but, when duplicated, a potential Sp1 site (with the consensus $5'^{G}/_{T}\ ^{G}/_{A}$ GGC$^{G}/_{T}\ ^{G}/_{A}\ ^{G}/_{A}\ ^{G}/_{T}$) and a potential AP2 site (with the consensus $5'$CCC$^{A}/_{C}$N$^{G}/_{C}\ ^{G}/_{C}\ ^{G}/_{C}$)[19] were introduced. However, functional analyses would be required in

order to determine whether these VNTR polymorphisms or base pair substitutions could lead to subtle variations in levels of expression.

Allele iii was the only variant which was common enough to investigate as a factor influencing the age of onset in HD. Although no relationship was detected, the power of our analysis was constrained by the small proportion of cases with allele iii and because the phase of this point mutation relative to the CAG repeat could not be determined, since we have investigated unrelated HD cases. Nevertheless, since all of the variants are rare, it is unlikely that they account for a substantial proportion of the general variance in age at onset of HD, which is not associated with the CAG sequence length.

The pattern of the VNTR polymorphisms of 6 and 20 bp in man suggest that these evolved by tandem duplication of the core sequence. Chimps and gorilla possess a single copy of both the 6 bp and the 20 bp stretch, and an allele with one repeat at each of these sites also occurs in man. Two simple models can explain our data. Either these loci are in the process of expanding in humans, or are on the point of contraction back to single copies and the process is complete in the apes. The former model is the most parsimonious, since these sequences are only present in single copies in chimps and gorilla. Thus it is likely that expansion has occurred in the human lineage, causing humans to be polymorphic at these sites. The commonest allele with one 6 bp and two 20 bp repeats is therefore not the ancestral allele. A precedent for this situation is seen, for example, in the apolipoprotein E gene in which the E3 allele occurs at frequencies of over 60%, while the less common E4 allele has been shown to be the ancestral state.[20]

The high degree of polymorphism in the 5' upstream region of the Huntington's gene highlights the need to establish the extent of normal genetic variability, particularly when studying disease associated genes and their control elements.

The first two authors contributed equally to this work. We would like to thank Mathias Chiano for carrying out statistical analyses, and Ronald Bontrop, David Craufurd, Alan Dodge, Gail Norbury, John Old, Elisabeth Rosser, and Katsushi Tokunaga for providing DNA samples. This work was supported by The Wellcome Trust (RC), The Rehabilitation and Medical Research Trust, and the Huntington's Disease Association, United Kingdom.

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* 1993;**72**:971-83.
2 Rubinsztein DC, Leggo J, Coles R, *et al.* Phenotypic characterization of individuals with 30-40 CAG repeats in the Huntington disease (HD) gene reveals HD cases with 36 repeats and apparently normal elderly individuals with 36-39 repeats. *Am J Hum Genet* 1996;**59**:16-22.
3 Nasir J, Goldberg YP, Hayden MR. Huntington disease: new insights into the relationship between CAG expansion and disease. *Hum Mol Genet* 1996;**5**:1431-5.
4 Rubinsztein DC, Barton DE, Davison BCC, Ferguson-Smith MA. 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 Mol Genet* 1993;**2**:1713-15.
5 Telenius H, Kremer HPH, Theilmann J, *et al.* Molecular analysis of juvenile Huntington disease: the major influence on $(CAG)_n$ repeat length is the sex of the affected parent. *Hum Mol Genet* 1993;**2**:1535-40.
6 Kremer B, Squitieri F, Telenius H, *et al.* Molecular analysis of late onset Huntington's disease. *J Med Genet* 1993;**30**: 991-5.
7 Burright EN, Clark HB, Servadio A, *et al.* SCA1 transgenic mice: a model for neurodegeneration caused by an expanded CAG trinucleotide repeat. *Cell* 1995;**82**:937-48.
8 Li SH, Schilling G, Young WS III, *et al.* Huntington's disease gene (IT15) is widely expressed in human and rat tissues. *Neuron* 1993;**11**:985-93.
9 Stine OC, Li SH, Pleasant N, Wagster MV, Hedreen JC, Ross CA. Expression of the mutant allele of IT-15 (the HD gene) in striatum and cortex of Huntington's disease patients. *Hum Mol Genet* 1995;**4**:15-18.
10 Lin B, Nasir J, Kalchman MA, *et al.* Structural analysis of the 5' region of mouse and human Huntington disease genes reveals conservation of putative promoter region and di- and trinucleotide polymorphisms. *Genomics* 1995;**25**: 707-15.
11 Vaishnaw AK, Hargreaves R, Campbell RD, Morley BJ, Walport MJ. Dnase I hypersensitivity mapping and promoter polymorphism analysis of human C4. *Immunogenetics* 1995;**41**:354-8.
12 Li WH, Sadler LA. Low nucleotide diversity in man. *Genetics* 1991;**129**:513-23.
13 Andersen LC, Beaty JS, Nettles JW, Seyfried CE, Nepom GT, Nepom RS. Allelic polymorphism in transcriptional regulatory regions of HLA-DQB genes. *J Exp Med* 1991;**173**:181-92.
14 Cereb N, Yang SY. The regulatory complex of HLA class I promoters exhibits locus-specific conservation with limited allelic variation. *J Immunol* 1994;**152**:3873-83.
15 Singal DP, Qiu X, D'Souza M, Sood SK. Polymorphism in the upstream regulatory regions of HLA-DRB genes. *Immunogenetics* 1993;**37**:143-7.
16 Yao Z, Volgger A, Scholz S, Albert ED. Sequence polymorphism in the HLA-B promoter region. *Immunogenetics* 1995;**41**:343-53.
17 Balbin M, Grubb A, Abrahamson M. Demonstration of sequence variations in the promoter region of the human cystatin C gene. *Biol Chem Hoppe-Seyler* 1992;**373**:471-6.
18 Fullerton SM, Harding RM, Boyce AJ, Clegg JB. Molecular and population genetic analysis of allelic sequence diversity at the human β-globin locus. *Proc Natl Acad Sci USA* 1994; **91**:1805-9.
19 Faisst S, Meyer S. Compilation of vertebrate-encoded transcription factors. *Nucleic Acids Res* 1992;**20**:3-26.
20 Hanlon CS, Rubinsztein DC. Arginine residues at codons 112 and 158 in the apolipoprotein E gene correspond to the ancestral state in humans. *Atherosclerosis* 1995;**112**:85-90.