Identification and characterization of the gene causing type 1 spinocerebellar ataxia

Sandro Banfi[1], Antonio Servadio[1], Ming-yi Chung[2], Thomas J. Kwiatkowski Jr.[3], Alanna E. McCall[1], Lisa A. Duvick[2], Ying Shen[3], Elizabeth J. Roth[1], Harry T. Orr[2] & H. Y. Zoghbi[1,3]

Spinocerebellar ataxia type 1 (SCA1) is a neurodegenerative disorder caused by expansion of a CAG trinucleotide repeat. In this study, we describe the identification and characterization of the gene harbouring this repeat. The *SCA1* transcript is 10,660 bases and is transcribed from both the wild type and SCA1 alleles. The CAG repeat, coding for a polyglutamine tract, lies within the coding region. The gene spans 450 kb of genomic DNA and is organized in nine exons. The first seven fall in the 5' untranslated region and the last two contain the coding region, and a 7,277 basepairs 3' untranslated region. The first four non-coding exons undergo alternative splicing in several tissues. These features suggest that the transcriptional and translational regulation of ataxin-1, the *SCA1* encoded protein, may be complex.

Departments of Pediatrics[1], Molecular and Human Genetics[3], Baylor College of Medicine, Houston, Texas 77030, USA [2]Departments of Laboratory Medicine and Pathology, and Biochemistry and Institute of Human Genetics, University of Minnesota, Minneapolis, Minnesota 55455, USA

Correspondence should be addressed to H.Y.Z.

Spinocerebellar ataxia type 1 (SCA1) is a dominantly inherited neurodegenerative disorder characterized by progressive neuronal loss in the cerebellum, brain stem and spinocerebellar tracts[1,2]. The main clinical features are ataxia, dysarthria, ophthalmoparesis and variable degrees of muscle wasting and neuropathy[3,4]. Usually symptoms develop in the third to fourth decade and the disease progressively worsens to cause death, due to bulbar dysfunction, 10 to 20 years after onset of symptoms. The juvenile onset with rapidly progressive course, reported in some families, suggests that anticipation occurs in this disease[5,6].

The gene that causes SCA1 has been localized to 6p22–p23 based on detailed genetic and physical mapping[7–11], and the mutation responsible for SCA1 has been determined to be an expansion of a CAG trinucleotide repeat[12]. The CAG repeat at the SCA1 locus is highly polymorphic and typically contains 6–39 repeat units on normal chromosomes whereas individuals with SCA1 have one allele in the normal size range and an expanded allele that contains 41–81 repeats[13–15]. Previous data revealed that the SCA1 CAG repeat is within an 11 kilobase (kb) transcript[12]. Limited sequence analysis of the region immediately flanking and containing the repeat suggested that the repeat is within the coding region of the gene and one of the possible open reading frames suggested that it encodes a polyglutamine tract[12]. In this study, we describe the identification and detailed characterization of the gene containing the unstable repeat at the SCA1 locus.

Isolation of *SCA1* cDNA

Two human fetal brain cDNA libraries were screened using various DNA fragments from the cosmid clone shown to contain the CAG repeat[12] (see Methodology) as probes. Five cDNA clones were identified; these included clone 31-5 containing the CAG repeat and clone 3J which was found not to overlap with 31-5 (Fig. 1). Northern blot analysis revealed that clones 31-5 and 3J identified the same 11 kb transcript detectable in all tissues examined (Fig. 2). Accordingly, the same two human fetal brain cDNA libraries and a human adult cerebellar cDNA library were used for several rounds of screening in order to obtain the full length transcript (see Methodology). As a result, 22 cDNA clones were isolated and characterized by sequence and PCR analyses to assemble a contig spanning the *SCA1* transcript. Twelve of the phage clones spanning the cDNA contig are shown in Fig. 1. We sequenced these clones and assembled the entire sequence of the *SCA1* cDNA which spans 10,660 bp (Fig. 3).

Sequence analysis revealed a coding region of 2,448 bp starting with an ATG codon at base 936 located within a nucleotide sequence that fulfills Kozak's criteria for an initiation codon[16]. An in-frame stop codon is present 57 bp upstream of this ATG in three independent cDNA clones as well as in genomic DNA. Furthermore, both the ATG at the beginning of the coding region and the upstream stop codon have been found in the murine homologue of *SCA1* (unpublished data). The *SCA1* gene is therefore predicted to encode an 87 kD protein, ataxin-1, which contains 816 amino acids, although one cannot exclude the possibility that the coding region begins at any of the other ATGs, located downstream of the first methionine, which would result in a smaller protein.

The CAG repeat is located within the coding region 588 bp from the first methionine and encodes a polyglutamine tract. The open reading frame ends with a TAG stop

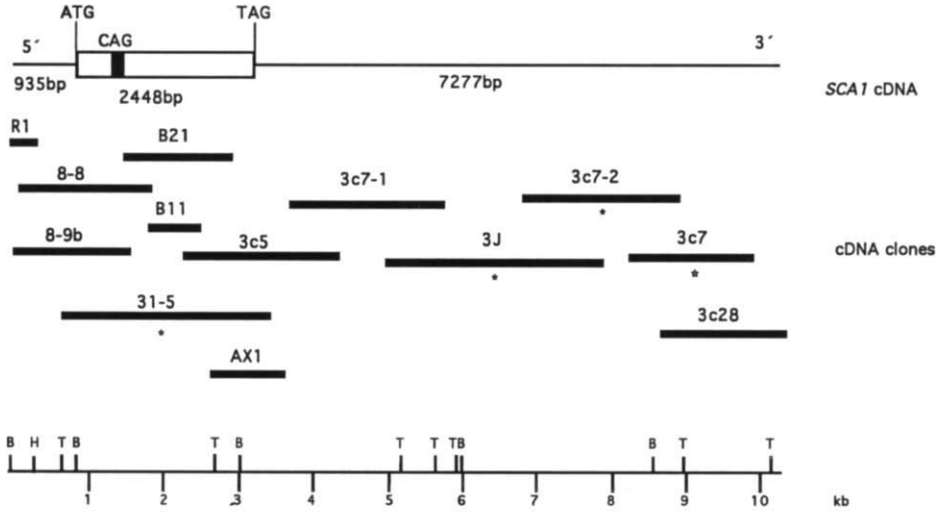

Fig. 1 Schematic representation of the *SCA1* cDNA contig. A subset of overlapping phage cDNA clones (black bars) and 5'-RACE-PCR product (R1) spanning 10.66 kb of the *SCA1* transcript is shown. The top scheme shows the structure of the *SCA1* transcript; the sizes of the coding region (rectangle) as well as the 5'UTR and the 3'UTR (thin lines) are indicated. The position of the CAG repeat within the coding region is also shown. An asterisk indicates the clones used as probes to screen the cDNA libraries. The positions of *Bam*HI (B), *Hin*dIII (H) and *Taq*I (T) restriction sites are shown at the bottom.

codon at base 3384. Therefore, this transcript has a 5' untranslated region (5'UTR) of 935 bases and a 3' untranslated region (3'UTR) of 7,277 bases. The transcript ends with a poly(A) tail of 57 residues; a polyadenylation signal, AATAAA, is found 23 nucleotides upstream of the poly(A) tail. Homology searches using both the DNA sequence of the coding region and the predicted protein sequence (lacking the CAG repeat and the polyglutamine tract, respectively) revealed no significant homology with other known proteins in the data base. Analysis of the sequence of ataxin-1 failed to reveal the presence of any strong phosphorylation sites nor any specific motifs such as DNA or RNA binding domains. The putative secondary structure of this protein is compatible with that of a soluble protein as no hydrophobic domains have been identified. A DNA sequence data base search revealed an identity between 380 bp in the 3'UTR of the *SCA1* transcript and an expressed sequence tag isolated from a human fetal brain cDNA library (EST04379)[17].

Alternative splicing in the 5'UTR

To characterize the genomic region flanking the CAG repeat, a 3.36 kb *Eco*RI genomic fragment known to contain this repeat[12] was completely sequenced. Alignment of this genomic sequence with the cDNA sequence allowed us to determine that the 3.36 kb *Eco*RI fragment contains a 2,080 bp exon which has 160 bp of 5'UTR, the putative initiation codon and the first 1,920 bp of the coding region. The rest of the coding region lies within the next downstream exon as detected by PCR analysis on genomic DNA. The last coding exon, which maps to a 9 kb *Eco*RI fragment in genomic DNA also contains 7,277 bp of 3'UTR for a total length of 7,805 bp (Fig. 4*a*).

Evidence for alternative splicing in the 5'UTR was initially suggested based on the hybridization pattern of the two most 5' cDNA clones, 8-8 and 8-9b (Fig. 1) to Southern blots containing *Eco*RI-digested genomic DNA from total human DNA and YACs spanning the SCA1 region. At least three strongly hybridizing fragments as well as the 3.36 kb *Eco*RI fragment were seen (data not shown). As neither of the cDNA clones contains an *Eco*RI site, this result suggested the presence of several exons in the 5'UTR of the *SCA1* transcript. Given these data and the unusual length of the 5'UTR, this region was characterized in more detail.

Alignment analysis of the sequence of clones 8-8 and 8-9b revealed the presence of two different 5' sequences diverging at basepair 322. This result strongly indicated of alternative splicing; to test this hypothesis, reverse transcription-PCR (RT-PCR) was performed on mRNA from cerebellar tissue using the primers shown in Fig. 3. When the primers 9b (specific for 8-9b clone) and 5R (present in both clones) were used in the RT-PCR analysis at least three products were obtained: one of the expected size (246 bp) and at least two fragments of larger sizes (see Fig. 4*b*). The same result was obtained when RT-PCR was carried out on liver, adrenal, brain and lymphoblast cDNAs (data not shown). The various RT-PCR products were cloned and sequenced. Sequence analysis of all these products and comparison with the sequence of phage clones 8-8 and 8-9b confirmed that they were the result of alternative splicing. Figure 4*a* shows the structure of all the cDNA clones that contain the 5' exons of the *SCA1* gene and depicts the various splice variants. Based on

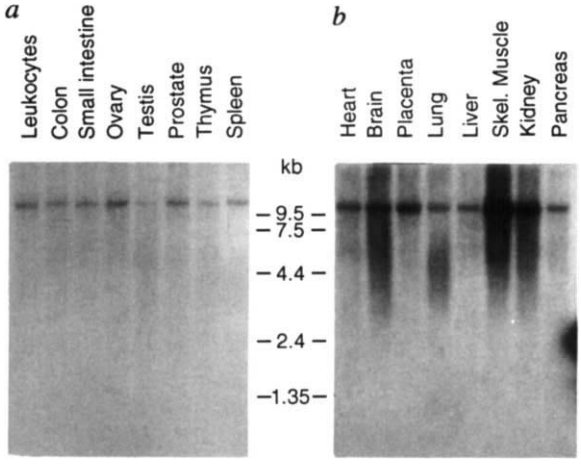

Fig. 2 Northern blot analysis of the *SCA1* gene using RNAs from multiple human tissues. *a*, RNA probed with PCR product from a portion of the coding region (bp 2460 to bp 3432). *b*, RNA probed with the 3J cDNA clone from the 3'UTR. An ~11 kb transcript is detected in RNA from all tissues using both probes as well as the cDNA clones 31-5 and 8-8, both of which contain the CAG repeat (Fig. 1).

sequence analysis of three cDNA clones and characterization of cerebellar RT-PCR products, five exons (1–5) were identified and their borders in the transcript were determined. Exons 2, 3 and 4 are alternatively spliced in the clones examined and in cerebellar tissue, whereas exon 5 was present in all the cDNA clones and RT-PCR products.

Rescreening of cDNA libraries with clones 8-8 and 8-9b as probes did not yield any additional cDNA clones. To identify other alternatively spliced exons in the 5'UTR

```
   1  CTACTACAGTGGCGGACGTACAGGACCTGTTTCACTGCAGGGGGATCCAAAACAAGCCCCGTGGAGCAACAGCCAGAGCAACAGCAGCTG    90
  91  CAAGACATTGTTTCTCTCCCTCTGCCCCCCCTTCCCCACGCAACCCCAGATCCATTTACACTTTACAGTTTTACCTCACAAAAACTACTA   180
 181  CAAGCACCAAGCTCCCTGATGGAAAGGAGCATCGTGCATCAAGTCACCAGGGTGGTCCATTCAAGCTGCAGATTTGTTTGTCATCCTTGT   270
 271  ACAGCAATCTCCTCCTCCACTGCCACTACAGGGAAGTGCATCACATGTCAGCATACTGGAGCATAGTGAAAGAGTCTATTTTGAAGCTTC   360
 361  AAACTTAGTGCTGCTGCAGACCAGGAACAAGAGAGAAAGAGTGGATTTCAGCCTGCACGGATGGTCTTGAAACACAAATGGTTTTTGGTC   450
 451  TAGGCGTTTTACACTGAGATTCTCCACTGCCACCCTTTCTACTCAAGCAAAATCTTCGTGAAAAGATCTGCTGCAAGGAACTGATAGCTT   540
 541  ATGGTTCTCCATTGTGATGAAAGCACATGGTACAGTTTTCCAAAGAAATTAGACCATTTTCTTCGTGAGAAAGAAATCGACGTGCTGTTT   630
 631  TCATAGGGTATTTCTCACTTCTCTGTGAAAGGAAGAAAGAACACGCCTGAGCCCAAGAGCCCTCAGGAGCCCTCCAGAGCCTGTGGGAAG   720
 721  TCTCCATGGTGAAGTATAGGCTGAGGCTACCTGTGAACAGTACGCAGTGAATGTTCATCCAGAGCTGCTGTTGGCGGATTGTACCCACGG   810
 811  GGAGATGATTCCTCATGAAGAGCCTGGATCCCCTACAGAAATCAAATGTGACTTTCCGTTTATCAGACTAAAATCAGAGCCATCCAGACA   900
 901  GTGAAACAGTCACCGTGGAGGGGGGACGGCGAAAAATGAAATCCAACCAAGAGCGGAGCAACGAATGCCTGCCTCCCAAGAAGCGCGAGA   990
   1                                  M  K  S  N  Q  E  R  S  N  E  C  L  P  P  K  K  R  E  I    19
 991  TCCCCGCCACCAGCCGGTCCTCCGAGGAGAAGGCCCCTACCCTGCCCAGCGACAACCACCGGGTGGAGGGCACAGCATGGCTCCCGGGCA  1080
  20    P  A  T  S  R  S  S  E  E  K  A  P  T  L  P  S  D  N  H  R  V  E  G  T  A  W  L  P  G  N    49
1081  ACCCTGGTGGCCGGGGCCACGGGGGCGGGAGGCATGGGCCGGCAGGGACCTCGGTGGAGCTTGGTTTACAACAGGGAATAGGTTTACACA  1170
  50    P  G  G  R  G  H  G  G  G  R  H  G  P  A  G  T  S  V  E  L  G  L  Q  Q  G  I  G  L  H  K    79
1171  AAGCATTGTCCACAGGGCTGGACTACTCCCCGCCCAGCGCTCCCAGGTCTGTCCCCGTGGCCACCACGCTGCCTGCCGCGTACGCCACCC  1260
  80    A  L  S  T  G  L  D  Y  S  P  P  S  A  P  R  S  V  P  V  A  T  T  L  P  A  A  Y  A  T  P   109
1261  CGCAGCCAGGGACCCCGGTGTCCCCCGTGCAGTACGCTCACCTGCCGCACACCTTCCAGTTCATTGGGTCCTCCCAATACAGTGGAACCT  1350
 110    Q  P  G  T  P  V  S  P  V  Q  Y  A  H  L  P  H  T  F  Q  F  I  G  S  S  Q  Y  S  G  T  Y   139
1351  ATGCCAGCTTCATCCCATCACAGCTGATCCCCCCAACCGCCAACCCCGTCACCAGTGCAGTGGCCTCGGCCGCAGGGGCCACCACTCCAT  1440
 140    A  S  F  I  P  S  Q  L  I  P  P  T  A  N  P  V  T  S  A  V  A  S  A  A  G  A  T  T  P  S   169
1441  CCCAGCGCTCCCAGCTGGAGGCCTATTCCACTCTGCTGGCCAACATGGGCAGTCTGAGCCAGACGCCGGGACACAAGGCTGAGCAGCAGC  1530
 170    Q  R  S  Q  L  E  A  Y  S  T  L  L  A  N  M  G  S  L  S  Q  T  P  G  H  K  A  E  Q  Q  Q   199
1531  AGCAGCAGCAGCAGCAGCAGCAGCAGCAGCATCAGCATCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCACCTCA  1620
 200    Q  Q  Q  Q  Q  Q  Q  Q  Q  H  Q  H  Q  Q  Q  Q  Q  Q  Q  Q  Q  Q  Q  Q  Q  Q  Q  H  L  S   229
1621  GCAGGGCTCCGGGGCTCATCACCCCGGGGTCCCCCCCACCAGCCCAGCAGAACCAGTACGTCCACATTTCCAGTTCTCCGCAGAACACCG  1710
 230    R  A  P  G  L  I  T  P  G  S  P  P  P  A  Q  Q  N  Q  Y  V  H  I  S  S  S  P  Q  N  T  G   259
1711  GCCGCACCGCCTCTCCTCCGGCCATCCCCGTCCACCTCCACCCCCACCAGACGATGATCCCACACACGCTCACCCTGGGGCCCCCCTCCC  1800
 260    R  T  A  S  P  P  A  I  P  V  H  L  H  P  H  Q  T  M  I  P  H  T  L  T  L  G  P  P  S  Q   289
1801  AGGTCGTCATGCAATACGCCGACTCCGGCAGCCACTTTGTCCCTCGGGAGGCCACCAAGAAAGCTGAGAGCAGCCGGCTGCAGCAGGCCA  1890
 290    V  V  M  Q  Y  A  D  S  G  S  H  F  V  P  R  E  A  T  K  K  A  E  S  S  R  L  Q  Q  A  I   319
1891  TCCAGGCCAAGGAGGTCCTGAACGGTGAGATGGAGAAGAGCCGGCGGTACGGGGCCCCGTCCTCAGCCGACCTGGGCCTGGGCAAGGCAG  1980
 320    Q  A  K  E  V  L  N  G  E  M  E  K  S  R  R  Y  G  A  P  S  S  A  D  L  G  L  G  K  A  G   349
1981  GCGGCAAGTCGGTTCCTCACCCGTACGAGTCCAGGCACGTGGTGGTCCACCCGAGCCCCTCAGACTACAGCAGTCGTGATCCTTCGGGGG  2070
 350    G  K  S  V  P  H  P  Y  E  S  R  H  V  V  V  H  P  S  P  S  D  Y  S  S  R  D  P  S  G  V   379
2071  TCCGGGCCTCTGTGATGGTCCTGCCCAACAGCAACACGCCCGCAGCTGACCTGGAGGTGCAACAGGCCACTCATCGTGAAGCCTCCCCTT  2160
 380    R  A  S  V  M  V  L  P  N  S  N  T  P  A  A  D  L  E  V  Q  Q  A  T  H  R  E  A  S  P  S   409
2161  CTACCCTCAACGACAAAAGTGGCCTGCATTTAGGGAAGCCTGGCCACCGGTCCTACGCGCTCTCACCCCACACGGTCATTCAGACCACAC  2250
 410    T  L  N  D  K  S  G  L  H  L  G  K  P  G  H  R  S  Y  A  L  S  P  H  T  V  I  Q  T  T  H   439
2251  ACAGTGCTTCAGAGCCACTCCCGGTGGGACTGCCAGCCACGGCCTTCTACGCAGGGACTCAACCCCCTGTCATCGGCTACCTGAGCGGCC  2340
 440    S  A  S  E  P  L  P  V  G  L  P  A  T  A  F  Y  A  G  T  Q  P  P  V  I  G  Y  L  S  G  Q   469
2341  AGCAGCAAGCAATCACCTACGCCGGCAGCCTGCCCCAGCACCTGGTGATCCCCGGCACACAGCCCCTGCTCATCCCGGTCGGCAGCACTG  2430
 470    Q  Q  A  I  T  Y  A  G  S  L  P  Q  H  L  V  I  P  G  T  Q  P  L  L  I  P  V  G  S  T  D   499
2431  ACATGGAAGCGTCGGGGGCAGCCCCGGCCATAGTCACGTCATCCCCCCAGTTTGCTGCAGTGCCTCACACGTTCGTCACCACCGCCCTTC  2520
 500    M  E  A  S  G  A  A  P  A  I  V  T  S  S  P  Q  F  A  A  V  P  H  T  F  V  T  T  A  L  P   529
2521  CCAAGAGCGAGAACTTCAACCCTGAGGCCCTGGTCACCCAGGCCGCCTACCCAGCCATGGTGCAGGCCCAGATCCACCTGCCTGTGGTGC  2610
 530    K  S  E  N  F  N  P  E  A  L  V  T  Q  A  A  Y  P  A  M  V  Q  A  Q  I  H  L  P  V  V  Q   559
2611  AGTCCGTGGCCTCCCCGGCGGCGGCTCCCCCTACGCTGCCTCCCTACTTCATGAAAGGCTCCATCATCCAGTTGGCCAACGGGGAGCTAA  2700
 560    S  V  A  S  P  A  A  A  P  P  T  L  P  P  Y  F  M  K  G  S  I  I  Q  L  A  N  G  E  L  K   589
2701  AGAAGGTGGAAGACTTAAAAACAGAAGATTTCATCCAGAGTGCAGAGATAAGCAACGACCTGAAGATCGACTCCAGCACCGTAGAGAGGA  2790
 590    K  V  E  D  L  K  T  E  D  F  I  Q  S  A  E  I  S  N  D  L  K  I  D  S  S  T  V  E  R  I   619
2791  TTGAAGACAGCCATAGCCCGGGCGTGGCCGTGATACAGTTCGCCGTCGGGGAGCACCGAGCCCAGGTCAGCGTTGAAGTTTTGGTAGAGT  2880
 620    E  D  S  H  S  P  G  V  A  V  I  Q  F  A  V  G  E  H  R  A  Q  V  S  V  E  V  L  V  E  Y   649
2881  ATCCTTTTTTTGTGTTTGGACAGGGCTGGTCATCCTGCTGTCCGGAGAGAACCAGCCAGCTCTTTGATTTGCCGTGTTCCAAACTCTCAG  2970
 650    P  F  F  V  F  G  Q  G  W  S  S  C  C  P  E  R  T  S  Q  L  F  D  L  P  C  S  K  L  S  V   679
2971  TTGGGGATGTCTGCATCTCGCTTACCCTCAAGAACCTGAAGAACGGCTCTGTTAAAAAGGGCCAGCCCGTGGATCCCGCCAGCGTCCTGC  3060
 680    G  D  V  C  I  S  L  T  L  K  N  L  K  N  G  S  V  K  K  G  Q  P  V  D  P  A  S  V  L  L   709
3061  TGAAGCACTCAAAGGCCGACGGCCTGGCGGGCAGCAGACACAGGTATGCCGAGCAGGAAAACGGAATCAACCAGGGGAGTGCCCAGATGC  3150
 710    K  H  S  K  A  D  G  L  A  G  S  R  H  R  Y  A  E  Q  E  N  G  I  N  Q  G  S  A  Q  M  L   739
3151  TCTCTGAGAATGGCGAACTGAAGTTTCCAGAGAAAATGGGATTGCCTGCAGCGCCCTTCCTCACCAAAATAGAACCCAGCAAGCCCGCGG  3240
 740    S  E  N  G  E  L  K  F  P  E  K  M  G  L  P  A  A  P  F  L  T  K  I  E  P  S  K  P  A  A   769
3241  CAACGAGGAAGAGGAGGTGGTCGGCGCCAGAGAGCCGCAAACTGGAGAAGTCAGAAGACGAACCACCTTTGACTCTTCCTAAGCCTTCTC  3330
 770    T  R  K  R  R  W  S  A  P  E  S  R  K  L  E  K  S  E  D  E  P  P  L  T  L  P  K  P  S  L   799
3331  TAATTCCTCAGGAGGTTAAGATTTGCATTGAAGGCCGGTCTAATGTAGGCAAGTAGAGGCAGCGTGGGGGAAAGGAAACGTGGCTCTCCC  3420
 800    I  P  Q  E  V  K  I  C  I  E  G  R  S  N  V  G  K  *                                       829
3421  TTATCATTTGTATCCAGATTACTGTACTGTAGGCTAAAATAACACAGTATTTACATGTTATCTTCTTAATTTTAGGTTTCTGTTCTAACC  3510
```

Fig. 3 The 5'UTR and the coding sequence of the *SCA1* transcript. The sequences of primers 9b, 5F and 5R (bases 129–147, bases 173–191 and bases 538–518 respectively in the 5' to 3' orientation) are underlined. The predicted protein sequence is shown below the DNA sequence. The full sequence including that of the 3'UTR (not shown here) is accessible through GenBank (X79204).

and to confirm our initial results, 5'-RACE-PCR[18] was carried out on reverse transcribed cerebellar mRNA using primers from the 5' end of exons 5 and 4. A 218 bp product was identified and its specificity was confirmed by Southern analysis using an internal PCR product as probe. Sequence analysis of the 5'-RACE-PCR product, furthermore, confirmed the alternative splicing of two exons (2 and 3) and allowed the identification of an additional 127 bp at the 5' end of this gene (Fig. 4*a*).

Intron–exon boundaries

Complete sequencing of the 3.36 kb *Eco*RI fragment provided the intron–exon boundaries for the 2,080 bp exon containing most of the coding region (Fig. 5). In order to determine the actual number of exons and to obtain all of the intron–exon boundaries, an inverse-PCR strategy was adopted using two overlapping YAC clones, 227B1 and 149H3 previously shown not to contain any rearrangements[11]. A total of nine exons, seven of which are in the 5'UTR, were identified and splice junctions for exons 1–9 were subcloned and sequenced (Fig. 5). The schematic diagram at the top of Fig. 4*a* shows these nine exons and their respective sizes. In the 5' untranslated region, alternative splicing involves exons 2, 3 and 4, but not exons 5, 6 and 7 in over 5 phage cDNA clones analysed. The putative exon 1 encompasses 157 bp and hybridizes very strongly to an *Eco*RI fragment derived from hamster genomic DNA (data not shown).

To study the genomic organization of the *SCA1* gene, ten cDNA clones and genomic fragments containing the splice junctions for all the exons were mapped by Southern analysis and localized on a long range restriction map of four overlapping YAC clones spanning the SCA1 critical region (Fig. 6). This analysis revealed that the gene spans at least 450 kb of genomic DNA and that the putative first exon maps to a genomic fragment containing a hypomethylated CpG island. Detailed restriction analysis of the intron between the two coding exons (8 and 9) revealed that this intron is approximately 4.5 kb in length. The sizes of the remaining introns were estimated from the long range restriction map and by PCR analysis and ranged from 650 bp (intron 2) to nearly 200 kb (intron 7) (Fig. 6).

Expression of the *SCA1* mRNA in patients

As a first step towards understanding the mechanism by which the expansion of a trinucleotide CAG repeat leads to neurodegeneration in SCA1, we examined *SCA1* transcription from the expanded alleles of patients. RT-PCR was carried out with primers Rep1 and Rep2 that flank the CAG repeat[12] using lymphoblastoid mRNAs from SCA1 patients with repeat sizes ranging from 43–69. This analysis revealed that mRNA was expressed from both the normal and the expanded alleles (Fig. 7).

Discussion

The mutation causing SCA1 has been identified as an expansion of a CAG repeat that is transcribed[12]. In this

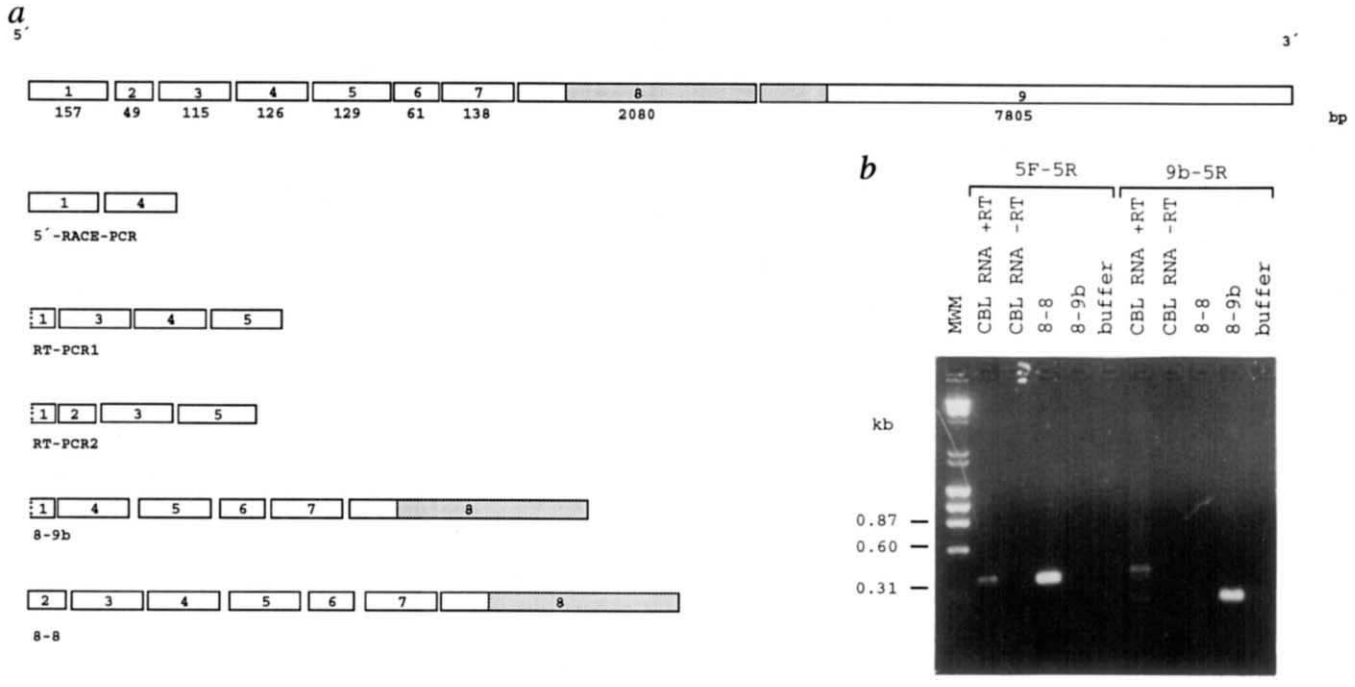

Fig. 4 *a*, The structure of the *SCA1* transcript and the various splice variants. The top scheme represents the nine exons (not drawn to scale) and their respective sizes. The stippled areas indicate the coding region (see Fig. 3). The structure of five cDNA clones representing different splice variants of the *SCA1* transcript are shown below. Clones 8-8 and 8-9b are phage clones, RT-PCR1 and RT-PCR2 are two clones obtained by RT-PCR carried out on cerebellar poly-(A)+ RNA using the primers 9b and 5R (Fig. 3). Only 30 bp of exon 1 were present in clone 8-9b and RT-PCR products as indicated by the dotted line in the rectangles. *b*, Detection of alternative splicing of the *SCA1* transcript in cerebellar poly-(A)$^+$ RNA (CBL RNA). RT-PCR analysis was carried out using two sets of primers: 9b-5R and 5F-5R. PCR products of the expected size were detected in CBL RNA in the presence of reverse transcriptase (+RT) with both pairs of primers. Using the 9b-5R pair at least two larger PCR products were also detected. Using the 5F-5R pair for RT-PCR and an annealing temperature of less than 60 °C some faint bands in the same size range as those seen using the 9b-5R primer pair were also seen (data not shown). 8-8 and 8-9b are the phage clones used as positive controls. The sizes of the relevant bands of the molecular weight marker (ΦX174 cut with *Hae*III) are indicated on the left.

		Exon 1	157 TTTACA	gtaagtga
158 gtttctatgcatag	GTTTTACC	Exon 2	206 GGAAAG	gtatatgg
207 ctcgaccattgcag	GAGCATCG	Exon 3	321 TGTCAG	gtgagagt
322 ttgtttgactgcag	CATACTGG	Exon 4	447 TTTTTG	gtaagtca
448 ttttataattacag	GTCTAGGC	Exon 5	575 GTACAG	gtaaacat
576 tttttctattccag	TTTTCCAA	Exon 6	637 CATAGG	gtgagtga
638 tatttccatgctag	GTATTTCT	Exon 7	775 AATGTT	gtaagtta
776 cttccctttcccag	CATCCAGA	Exon 8	2855 GCCCAG	gtaacgtt
2857 ccctgtttccacag	GTCAGCGT	Exon 9		
YYYYYYYYYYNCAG		Consensus	AG	GTRAGT

Fig. 5 Intron–exon boundaries of the *SCA1* gene. Splice acceptor and splice donor sites are indicated in bold letters. The numbers at the beginning and the end of each exon refer to the position in the composite sequence of *SCA1* in Fig. 3. Uppercase letters indicate exon sequences, lowercase letters indicate intron sequences. Y, pyrimidine; R, purine; N, undefined.

study, we report the identification and characterization of the gene involved in SCA1. The *SCA1* transcript is estimated to be 10.5–11 kb based on northern analysis. It has a wide pattern of expression which includes both neuronal and non-neuronal tissues. Sequence analysis of several overlapping cDNA clones has allowed us to assemble 10,660 bp of sequence and identify a coding region of 2,448 bp.

The *SCA1* transcript has 935 bases in the 5'UTR and a large 3'UTR of 7,277 bases which is currently the longest 3'UTR in Genbank. The gene locus spans approximately 450 kb and is organized in nine exons. The organization of *SCA1* is unusual when compared to most mammalian genes. The coding region falls within two large exons measuring 2,080 bp and 7,805 bp respectively. The remaining seven exons represent the 5'UTR and, unlike the coding exons, these are small (49–157 bp) and separated by large introns to encompass 400 kb of genomic DNA. Exons 2–4 undergo alternative splicing which does not vary in the five different tissues examined. The large number of non coding exons led us to characterize several cDNA clones and portions of the genomic locus to examine whether there are any significant open reading frames 5' to the first putative methionine. The putative initiation codon at basepair 936, has been found in four independent cDNA clones from three different libraries, in genomic clones, and in the murine homologue (unpublished data). Furthermore a stop codon, 19 amino acids 5' of the putative initiation codon was present in cDNA and in genomic clones as well as in the mouse homologue. This stop codon is in exon eight, the first of the two coding exons. The possibility that some of the 5' exons are translated to generate an isoform of ataxin-1 is unlikely given that there are stop codons in all three reading frames in exons 4–8. No methionine in the correct reading frame was detected in exons 1 and 2. Even if a methionine was to be identified 5' to the available sequence, exons 1–3 have to be spliced to a sequence within the coding region, past the stop codons in exons 4–8. Such a splice variant would generate a smaller transcript likely to be detected by northern analysis. However, there are eleven ATGs 5' to the putative initiation codon and some of these fall within acceptable Kozak consensus sequences. It is therefore possible that these ATGs might play a role in regulation of translation[16].

Mammalian genes with either large internal exons or large 5'UTR are uncommon. Examples of genes with large internal exons include coagulation factors V and VIII and apolipoprotein B which have internal exons measuring 2820, 3106 and 7572, respectively[19,20]. Human genes with a large 5'UTR include the acidic fibroblast growth factor and c-*fgr* proto-oncogene which have four and seven 5' untranslated exons respectively[21,22]. The 5' non coding exons in the c-*fgr* gene undergo extensive alternative splicing as was observed in *SCA1* (ref. 21). The function of the large 5'UTR in *SCA1* is not clear at this point. Given that exon 1 appears to be conserved in hamster, based on hybridization data, and that exon 7 is conserved in mouse, based on sequence analysis (unpublished data), it is possible that the 5' UTR of the *SCA1* gene has a regulatory function at the transcriptional or translational level.

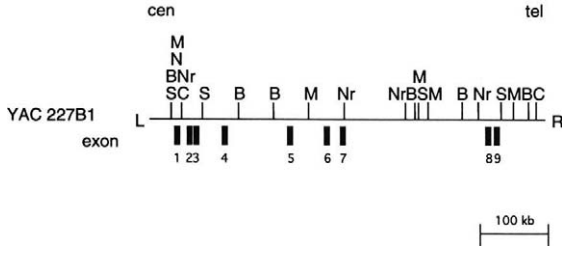

Fig. 6 Genomic structure of the *SCA1* gene. The nine exons of the *SCA1* gene (solid rectangles not drawn to scale) were localized based on the restriction map of the SCA1 region by Southern analysis using rare cutter DNA digests from several YAC clones. A representative map using YAC clone 227B1, which encompasses the *SCA1* gene, is shown. The restriction map of this YAC has been confirmed by analysis of four overlapping YAC clones in the region. The centromere–telomere orientation is indicated by Cen and Tel. L, left YAC end; R, right YAC end; B, *Bss*HII; C, *Csp*I; M, *Mlu*I; N, *Not*I; Nr, *Nru*I; S, *Sac*II.

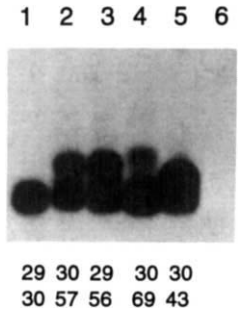

Fig. 7 Analysis of expression of the expanded *SCA1* allele. RT-PCR was carried out on lymphoblast poly-(A)⁺RNA from one unaffected individual (lane 1) and four SCA1 patients (lanes 2–5) using primers Rep1 and Rep2. This analysis shows that both the normal and the expanded *SCA1* alleles are transcribed. The number of the repeat units for each allele is indicated below each lane; lane 6 is the RT minus control.

Isolation and characterizaton of the first six exons from the murine homologue is currently in progress and may identify important regulatory sequences. Alternatively, sequences in the 5' or 3' UTR of *SCA1* may have specific biologic activity, as seen for instance in a study by Rastinejad *et al.*[23] which demonstrated that a 200 bp segment within the 3' UTR of the *α-tropomyosin* gene suppresses tumour formation in mouse muscle.

SCA1 normally encodes a predicted protein of 792–825 amino acids where the variability in size correlates with the length of the polymorphic CAG repeat encoding a polyglutamine stretch. Data base searches did not identify any homology between this predicted protein and previously identified molecules. The mechanism by which the polyglutamine tract expansion leads to specific Purkinje cell loss in SCA1 is not clear. Expansion of a trinucleotide repeat coding a polyglutamine tract has proved to be the mutational mechanism for three other neurodegenerative disorders, spinal bulbar muscular atrophy (SBMA)[24], Huntington's disease (HD)[25] and hereditary dentatorubralpallidoluysian atrophy (DRPLA)[26,27]. It is interesting to note some similarities between the genes involved in these diseases. The polyglutamine tract is located near the amino terminus of the protein in SCA1, HD and SBMA. The androgen receptor, involved in SBMA, shares similar features with *SCA1*. These include a very long 5'UTR (1,126 bp) and a very large internal exon (2,712 bp)[28] harbouring the unstable CAG repeat. A long 3'UTR is also present in the *HD* mRNA (up to 3,921 bp)[29]. In each of these diseases the gene has a wide pattern of expression but the neurodegeneration involves a specific cell type. The mechanism by which this molecular event leads to neurodegeneration is still unknown and is, very likely, common to all the above diseases. Expanded trinucleotide repeats have been proven to interfere with transcription in fragile X syndrome[30] but in myotonic dystrophy the effect of the expansion on transcription is not yet clear[31,32]. Therefore, we studied the expression of the *SCA1* gene in patients with SCA1. This analysis revealed that the expanded CAG repeat does not result in the elimination of the *SCA1* transcription as transcripts from the expanded alleles on SCA1 chromosomes were detected in lymphoblasts. This result suggests that impairment of transcription efficiency is not the mechanism underlying the neurodegeneration in SCA1.

The expansion of the polyglutamine tracts in SBMA, HD, SCA1 and DRPLA most likely leads to a gain of function or to a dominant negative effect since HD, SCA1 and DRPLA are dominant and deletions of the androgen receptor[33,34] or hemizygosity at the HD locus[25] do not cause SBMA and HD, respectively. We propose that, in these disorders, a protein with an expanded polyglutamine tract retains its normal function totally or partially (as in the case of the androgen receptor in SBMA) but a gain of function or dominant negative effect results from an aberrant interaction with itself, its normal target or a new gene product. This aberrant interaction would eventually lead to the cell-specific neurodegenerative process seen in these four diseases. The characterization of the *SCA1* transcript, and identification of the coding portion of this molecule is an important step towards studies aimed at characterizing the ataxin-1 protein and pathogenetic mechanism in SCA1.

Methodology

Screening of cDNA libraries. Three cDNA libraries were screened: a human fetal brain library from Stratagene, a human fetal brain

Table 1 Primer sequences for inverse-PCR

Exon	Primer 1	Primer 2
2	X2-1 GTAGTAGTTTTTGTGAGG (181–164)	X2-2 CACCAAGCTCCCTGATGGA (185–203)
3	X3-1 GCTTGAATGGACCACCCT (246–229)	X3-2 ATCTCCTCCTCCACTGCCAC (277–296)
4	X4-1 AGACTCTTTCACTATGCTC (347–329)	X4-2 TTCAGCCTGCACGGATGGT (407–425)
5	5a TGGCAGTGGAGAATCTCAGT (482–463)	5-2 TGCTGCAAGGAACTGATAGC (519–538)
6	10a AATGGTCTAATTTCTTTGG (598–580)	10b GAGAAAGAAATCGACGTGC (607–625)
7	6-1 ACAGGCTCTGGAGGGCTCCT (714–695)	X5-2 TCCATGGTGAAGTATAGGCT (723–742)
9	9-1 AGCAGGATGACCAGCCCTGT (2919–2900)	9-2 GCTCTTTGATTTGCCGTGT (2939–2957)

All primers are read in the 5' to 3' direction. Numbers in parentheses represent the coordinates of each primer within the SCA1 cDNA sequence Fig. 3).

library constructed in λ-Zap II (kindly provided by Cheng Chi Lee, Baylor College of Medicine) and an adult cerebellar cDNA library from Clonetech. Libraries were plated on 150 cm plates at a density of 50,000 pfu per plate using bacterial strain LE392. Hybond-N filters (Amersham) were used to carry out plaque lifts. The fragments used as probes in the first screening included a mixture of two polymerase chain reaction (PCR) products obtained by using the primers Rep1 and Rep2 (ref. 12) immediately flanking the repeat and the primers Pre1 and Pre2 (ref. 12) which amplify a sequence immediately adjacent to the repeat, and a 1.1 kb subclone of the 3.36 kb *Eco*RI fragment[12]. The 1.1 kb fragment is located 540 bp 3' to the CAG repeat. A 9 kb *Eco*RI genomic fragment derived from the same cosmids containing the CAG repeat was also used in this screening. Subsequent rounds of screening were carried out on the same libraries using as probes cDNA clones 31-5, 3J, 3c7-2 and 3c7 (Fig. 1). Genomic and cDNA probes were labelled using the random priming technique[35]; repetitive sequences were blocked using sheared human placental DNA as described previously[36]. Hybridization of the filters was then carried out following standard protocols[37].

DNA sequencing and sequence analysis. Shotgun libraries were constructed in M13 (ref. 38) for each of the following cDNA clones: 8-8, 31-5, 3c5, 3c7-1, 3J, 3c7-2 and 3c7 (Fig. 1). 20–30 M13 subclones were sequenced for each cDNA clone using an Applied Biosystem, ABI 370A, automated fluorescent sequencer, as described[39]. Some cDNA clones (8-9b, 8-9a, AX1, B21, B11, 3c28) have been partially sequenced manually using a Sequenase sequencing kit (USB) on double-stranded templates. The sequence coverage in terms of numbers of cDNA/genomic clones analysed was 3–4X in the coding and 5'UTR and 2X in the 3'UTR. All the RT-PCR, the 5'-RACE-PCR and the inverse-PCR products were sequenced manually after subcloning into *Sma*I-digested pBluescript SK- plasmid (Stratagene) modified using the T-vector protocol[40].

Data base searches were carried out using the GCG software package and the BLAST network service from the National Center for Biotechnology Information[41]. The sequence of the *SCA1* transcript has been deposited in Genbank, accession number X79204.

Northern blot, RT-PCR and genomic PCR analyses. The northern blot of poly-$(A)^+$ RNA from various human tissues and the poly-$(A)^+$ RNA from adult human cerebellum were purchased from Clonetech. Poly-$(A)^+$ RNA from human lymphoblastoid cells was prepared as described[12]. First strand randomly primed cDNA synthesis was carried out using MMLV reverse transcriptase (BRL). (First strand randomly primed cDNA from human brain, liver and adrenal were provided to us by G. Borsani, Baylor College of Medicine). RT-PCR for detection of alternative splicing was carried out with primers 9b and 5R and with primers 5F and 5R (Fig. 3) under the following conditions: initial denaturation step at 94 °C for 5 min followed by 30 cycles of 94 °C for 1 min, 60 °C for 1 min and 72 °C for 2 min.

RT-PCR on lymphoblastoid cell lines with primers Rep1 and Rep2 for detection of expression of ataxin-1 from SCA1 chromosomes was performed as previously described[12]. 20 μl of the PCR reactions was then resolved on a 2% agarose gel and blotted onto Sureblot membrane (Oncor). The filter was hybridized with a $(GCT)_7$ oligonucleotide end-labelled with γ-^{32}P-ATP according to our published protocol[12].

5'-RACE-PCR. First strand cDNA was prepared from 1 μg of poly-$(A)^+$ RNA from human adult cerebellum (Clonetech) using the primer 5R (Fig. 3). 5'-RACE-PCR was carried out as previously described[18] using *SCA1* primers 5a and X4-1 (Table 1) as specific primers. The product was electrophoresed through a 1.2 % agarose gel, blotted onto SureBlot hybridization membrane (Oncor) and, to test the specificity of the product, hybridized to a *SCA1* specific probe represented by a PCR product spanning 118 bp between primer 9b in exon 1 and primer X3-1 (Table 1) in exon 3.

Identification of intron–exon boundaries. The boundaries of exons 2–9 were identified by inverse-PCR. To carry out inverse-PCR, YAC plugs were digested to completion as described[42] using frequent-cutter restriction enzymes such as *Sau*3aI, *Taq*I , *Hae*III and *Msp*I (Boehringer Mannheim Biochemicals) and used as recommended by the manufacturer. The plugs were digested with β agarase I (USB) following the manufacturer's recommendations and subsequently phenol-chloroform extracted, precipitated with ethanol and resuspended in 12 μl of TE pH 8. 50 ng of DNA from each digest was circularized according to the published protocol[43]. Diverging PCR primers were designed within the cDNA (Table 1) and used on the above circularized product under the conditions described by Groden *et al.*[43]. PCR products were then subcloned and sequenced. Inverse-PCR failed to identify the boundary of exon 1. Accordingly, a 9 kb *Eco*RI genomic fragment found to contain exon 1 was subcloned from a cosmid derived from YAC 227B1. This subclone was subsequently partially sequenced to identify the boundary of exon 1.

Mapping cDNA clones to the YACs and cosmids. Southern blots containing *Eco*RI-digested DNAs from YACs spanning the SCA1 critical region[11] as well as Southern blots containing DNAs from the YACs digested with rare-cutter enzymes[11] were hybridized, using our standard protocols[37], to various *SCA1* cDNA clones and to all the genomic fragments containing the intron-exon boundaries.

Received 6 April; accepted 17 May 1994.

Acknowledgements

We thank A. Ballabio, D. Nelson and P. Patel for the critical reading of the manuscript; and R. Gibbs and D. Muzny for helpful discussions. This work was supported by grants from the National Institutes of Health (NS27699 and NS22920) and the Muscular Dystrophy Association. Portions of this work were supported by the core facilities of the MRRC and the Human Genome Center (Baylor College of Medicine).

1. Greenfield, J.G. *The spino-cerebellar degenerations* (Charles C. Thomas, Springfield, Illinois, 1954).
2. Zoghbi, H.Y. The spinocerebellar degenerations in *Current Neurology* (ed. Appel, S.H.) 121–144 (Mosby-Year Book, St-Louis, 1991).
3. Schut, J.W. Hereditary ataxia: clinical study through six generations. *Arch. Neurol. Psychiat.* **63**, 535–567 (1954).
4. Currier, R.D., Glover, G., Jackson, J.F. & Tipton, A.C. Spinocerebellar ataxia: study of a large kindred. *Neurology* **22**, 1040–1043 (1972).
5. Haines, J.L., Schut, L.J. & Weitkamp, L.R. Spinocerebellar ataxia in large kindred: age at onset, reproduction, and genetic linkage studies. *Neurology* **34**, 1542–1548 (1984).
6. Zoghbi, H.Y. *et al.* Spinocerebellar ataxia: variable age of onset and linkage to human leukocyte antigen in a large kindred. *Ann. Neurol.* **23**, 580–584 (1988).
7. Jackson, J.F., Currier, R.D., Terasaki, P.I. & Morton, N.E. Spinocerebellar ataxia and HLA linkage: risk prediction by HLA typing. *New Engl. J. Med.* **296**, 1138–1141 (1977).
8. Zoghbi, H.Y. *et al.* The gene for autosomal dominant spinocerebellar ataxia (SCA1) maps telomeric to HLA complex and is closely linked to the D6S89 locus in three large kindreds. *Am. J. hum. Genet.* **49**, 23–30 (1991).
9. Ranum, L.P.W. *et al.* Localization of the autosomal dominant, HLA-linked spinocerebellar ataxia (SCA1) locus in two kindreds within an 8cM subregion of chromosome 6p. *Am. J. hum. Genet.* **49**, 31–41 (1991).
10. Kwiatkowski Jr, T.J. *et al.* The gene for autosomal dominant spinocerebellar ataxia (SCA1) maps centromeric to D6S89 and shows no recombination, in nine large kindreds, with a dinucleotide repeat at the AM10 locus. *Am. J. hum. Genet.* **53**, 391–400 (1993).
11. Banfi, S. *et al.* Mapping and cloning of the critical region for the spinocerebellar ataxia type 1 gene in a yeast artificial chromosome contig spanning 1.2Mb. *Genomics* **18**, 627–635 (1993).
12. Orr, H. *et al.* Expansion of an unstable trinucleotide (CAG) repeat in spinocerebellar ataxia type 1. *Nature Genet.* **4**, 221–226 (1993).
13. Matilla, T. *et al.* Presymptomatic analysis of spinocerebellar ataxia type 1 (SCA1) via the expansion of the *SCA1* CAG-repeat in a large pedigree displaying anticipation and parental male bias. *Hum. molec. Genet.* **2**, 2123–2128 (1993).
14. Jodice, C. *et al.* Effect of trinucleotide repeat length and parental sex on phenotypic variation in spinocerebellar ataxia 1. *Am. J. hum. Genet.* **54**, 959–965 (1994).
15. Ranum, L.P.W. *et al.* Molecular and clinical correlations in spinocerebellar ataxia type 1 (SCA1): evidence for familial effects on the age of onset. *Am. J. hum. Genet.* **55**, 244–252 (1994).
16. Kozak, M. The scanning model for translation: an update. *J. cell. Biol* **108**, 229–241 (1989).
17. Adams, M.D., Kerlavage, A.R., Fields, C. & Venter, J.C. 3400 Expressed sequence tags identity diversity of transcripts from human brain. *Nature Genet.* **4**, 256–267 (1993).
18. Frohman, M.A. Race: rapid amplification of cDNA ends in *PCR protocols. A guide to methods and applications* (eds Innis, M.A., Gelfand, D.H., Sninsky, J.J. & Whit, T.J.) (Academic Press, San Diego, 1990).
19. Cripe, L.D., Moore, K.D. & Kane, W.H. Structure of the gene for human coagulation factor V. *Biochem.* **31**, 3777–3785 (1992).
20. Ludwig, E.H. *et al.* DNA sequence of the human apolipoprotein B gene. *DNA* **6**, 363–372 (1987).
21. Myers, R.L., Payson, R.A., Chotani, M.A., Deaven, L.L. & Chiu, I.M. Gene structure and differential expression of acidic fibroblast growth factor mRNA: identification and distribution of four different transcripts. *Oncogene* **8**, 341–349 (1993).
22. Link, D.C., Gutkind, S.J., Robbins, K.C. & Ley, T.J. Characterization of the 5' region of the human c-fgr and identification of the major myelomonocytic c-fgr promoter. *Oncogene* **7**, 877–884 (1992).
23. Rastinejad, F., Conboy, M.J., Rando, T.A. & Blau, H.M. Tumor Suppression by RNA from the 3' untranslated region of α-tropomyosin. *Cell* **75**, 1107–1117 (1993).
24. La Spada, A.R., Wilson, E.M., Lubahn, D.B., Harding, A.E. & Fischbeck, H. Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. *Nature* **352**, 77–79 (1991).
25. 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).
26. Koide, R. *et al.* Unstable expansion of CAG repeat in hereditary dentatorubral-pallidoluysian atrophy (DRPLA). *Nature Genet.* **6**, 9–13 (1994).
27. Nagafuchi, S. *et al.* Dentatorubral and pallidoluysian atrophy expansion of an unstable CAG trinucleotide on chromosome 12p. *Nature Genet.* **6**, 14–18 (1994).
28. Tilley, W.D., Marcelli, M. & McPhaul, M.J. Expression of the human androgen receptor gene utilizes a common promoter in diverse human tissues and cell lines. *J. biol. Chem.* **265**, 13776–13781 (1990).
29. Lin, B. *et al.* Differential 3' polyadenylation of the Huntington disease gene results in two mRNA species with variable tissue expression. *Hum. molec. Genet.* **2**, 1541–1545 (1993).
30. Pieretti, M. *et al.* Absence of expression of the *FMR-1* gene in fragile X syndrome. *Cell* **66**, 817–822 (1991).
31. Fu, Y.-H. *et al.* Decreased expression of myotonin-protein kinase messenger RNA and protein in adult form of myotonic dystrophy. *Science* **260**, 235–238 (1993).
32. Sabouri, L.A. *et al.* Effect of the myotonic dystrophy (DM) mutation on mRNA levels of the DM gene. *Nature Genet.* **4**, 233–238 (1993).
33. Trifiro, M. *et al.* The 56/58 kDa androgen-binding protein in male genital skin fibroblasts with a deleted androgen receptor gene. *Molec. cell. Endocrinol.* **75**, 37–47 (1991).
34. Quigley, C.A. *et al.* Complete deletion of the androgen receptor gene: definition of the null phenotype of the androgen insensitivity syndrome and determination of carrier status. *J. clin. Endocrinol. Metab.* **74**, 927–933 (1992).
35. Feinberg, A.P. & Vogelstein, B. A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. *Analytical Biochem.* **137**, 266–267 (1984).
36. Sealy, P.G., Whittaker, P.A. & Southern, E.M. Removal of repeated sequences from hybridization probes. *Nucl. Acids Res.* **13**, 1905–1922 (1985).
37. Zoghbi, H.Y., Daiger, S.P., McCall, A., O'Brien, W.E. & Beaudet, A.L. Extensive DNA polymorphism at the factor XIIIa (F13a) locus and linkage to HLA. *Am. J. hum. Genet.* **42**, 877–883 (1988).
38. Bankier, A.T., Weston, K.M. & Barrel, B.G. Random cloning and sequencing by the M13 dideoxynucleotide termination method. *Meth. Enzymol.* **155**, 55–93 (1987).
39. Gibbs, R., Nguyen, P.N., Mc Bride, L.J., Koepf, S.M. & Caskey, C.T. Identification of mutations leading to the Lesch-Nyhan syndrome by automated direct DNA sequencing of *in vitro* amplified cDNA. *Proc. natn. Acad. Sci. U.S.A.* **86**, 1919–1923 (1989).
40. Marchuk, D., Drumm, M., Saulino, A. & Collins, F.S. Construction of T-vectors, a rapid and general system for direct cloning of unmodified PCR products. *Nucl. Acids Res.* **19**, 1154 (1990).
41. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. A basic local alignment search tool. *J. molec. Biol.* **215**, 403–410 (1990).
42. Wapenaar, M.C. *et al.* The genes for X-linked ocular albinism (OA1) and microphthalmia with linear skin defects (MLS): cloning and characterization of the critical regions. *Hum. molec. Genet.* **2**, 947–952 (1993).
43. Groden, J. *et al.* Identification and characterization of the familial adenomatous polyposis Coli gene. *Cell* **66**, 589–600 (1991).