Moderate expansion of a normally biallelic trinucleotide repeat in spinocerebellar ataxia type 2

Stefan-M. Pulst[1], Alex Nechiporuk[1*], Tamilla Nechiporuk[1*], Suzana Gispert[2], Xiao-Ning Chen[7], Iscia Lopes-Cendes[6], Susan Pearlman[4], Sidney Starkman[4], Guillermo Orozco-Diaz[5], Astrid Lunkes[2], Pieter DeJong[3], Guy A. Rouleau[6], Georg Auburger[2], Julie R. Korenberg[7], Carla Figueroa[1] & Soodabeh Sahba[1]

[1]*The Rose Moss Laboratory for Parkinson's and Neurodegenerative Diseases, CSMC Burns and Allen Research Institute, and Division of Neurology, Cedars-Sinai Medical Center, UCLA School of Medicine, Los Angeles, California 90048, USA*
[2]*Department of Neurology, Heinrich-Heine Universitaet, 4000 Duesseldorf 1, Germany,*
[3]*Department of Human Genetics, Roswell Park Cancer Institute, Buffalo, New York 14263-0001, USA*
[4]*Department of Neurology, UCLA School of Medicine, Los Angeles, CA 90069, USA*
[5]*Neurology Service, Lenin Hospital, Holguin, Cuba*
[6]*Centre for Research in Neuroscience, The Montreal General Hospital, McGill University, Montreal, Quebec, Canada*
[7]*Division of Medical Genetics, Cedars-Sinai Medical Center, UCLA School of Medicine, Los Angeles, California 90048, USA*

[*]*A.N. & T.N. contributed equally to the project*

Correspondence should be addressed to S.M.P.

The gene for spinocerebellar ataxia type 2 (*SCA2*) has been mapped to 12q24.1. A 1.1-megabase contig in the candidate region was assembled in P1 artificial chromosome and bacterial artificial chromosome clones. Using this contig, we identified a CAG trinucleotide repeat with CAA interruptions that was expanded in patients with SCA2. In contrast to other unstable trinucleotide repeats, this CAG repeat was not highly polymorphic in normal individuals. In SCA2 patients, the repeat was perfect and expanded to 36–52 repeats. The most common disease allele contained $(CAG)_{37}$, one of the shortest expansions seen in a CAG expansion syndrome. The repeat occurs in the 5'-coding region of *SCA2* which is a member of a novel gene family.

The hereditary ataxias are a complex group of neurodegenerative disorders all characterized by varying abnormalities of balance attributed to dysfunction or pathology of the cerebellum and cerebellar pathways. In many of these disorders, dysfunction or structural abnormalities extend beyond the cerebellum, and may involve basal ganglia function, oculo-motor disorders and neuropathy. The dominant spinocerebellar ataxias (SCAs) represent a phenotypically heterogeneous group of disorders with a prevalence of familial cases of approximately 1 in 100,000 (ref. 1).

The genes causing two types of SCA have recently been identified: *SCA1* on chromosome 6p (ref. 2) and the gene for Machado-Joseph disease (MJD) on chromosome 14q (ref. 3). These diseases are caused by expansion of a CAG repeat in the coding region of the genes. However, many SCA pedigrees did not show linkage to chromosome 6p or 14q, confirming the presence of non-allelic heterogeneity. Subsequent genetic linkage studies have led to the identification of loci for *SCA2* on chromosome 12 (refs 4,5), *SCA4* on chromosome 16 (ref. 6), *SCA5* on chromosome 11 (ref. 7) and *SCA7* on chromosome 3 (ref. 8).

The location of *SCA2* on human chromosome 12 has recently been refined to a 1-cM interval between *D12S1328* and *D12S1333* or *D12S1329* in a large Cuban pedigree[9]. A similar location of *SCA2* has been suggested for pedigrees of German, French-Canadian, Tunisian, and Italian origin[11]. One marker of unknown physical location for the locus *D12S1332* did not detect the recombinants detected by *D12S1328* or *D12S1333* and likely represented the genetic marker closest to *SCA2*. The physical map of the *SCA2* region on chromosome 12q24.1 contains at least two gaps and *D12S1333* could not be placed on the map[10,11]. In addition, discrepancies between the physical and genetic map could not be resolved, because a YAC contig spanning the region bounded by the genetic flanking markers could not be assembled.

Anticipation has been observed in SCA2[5,12] suggesting that SCA2 like other neurodegenerative disorders such as Huntington disease (HD), SCA1 and SCA3, spinobulbar muscular atrophy (SBMA) and dentato-rubropallido-luysian atrophy (DRPLA) was caused by unstable DNA repeats[2,3,13-16]. This notion received further support with the identification of a 150-kD protein in a patient with chromosome 12-linked ataxia using a monoclonal antibody that detects extended polyglutamine tracts[17]. We now report the identification of the *SCA2* gene.

Physical map of the *SCA2* region

Due to the lack of a YAC contig spanning the *SCA2* candidate region we generated a contig of P1-artificial-chromosomes (PACs)[18] and bacterial artificial chromosome (BACs) (Fig. 1). Additional contigs were established using a FISH-mapped BAC resource (J.R.K., unpublished observations) by selecting clones mapping to the proximal third of band 12q24.1. After several walking steps using STSs derived from PAC- and BAC- end sequences[19], overlap of the contigs was established by shared STS content. Overlap of clones was further confirmed by shared restriction fragments through hybridization of selected clones to Southern blots of *Not*I/*Xba*I digests of clones (data not shown). All STSs were mapped back to human chromosome 12 by PCR

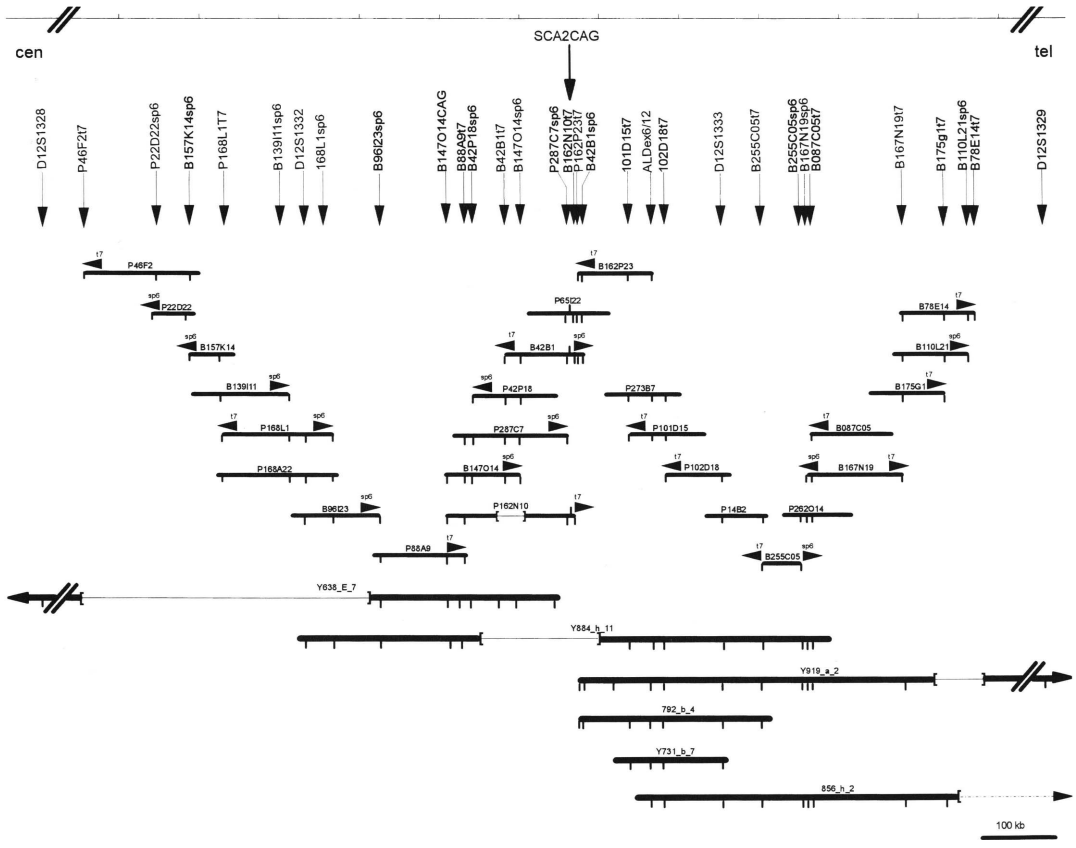

Fig. 1 Physical map of the *SCA2* region. The location of *D12S1328* centromeric and *D12S1329* telomeric of the contig are indicated. As indicated by double forward slashes, the map is not drawn to scale between *D12S1328* and *P46F2t7*, and between *B78E14t7* and *D12S1329*. YAC, PAC and BAC clones are prefixed with 'Y', 'P', and 'B' respectively. Clones positive for a specific STS by PCR analysis are indicated by vertical lines. Solid arrows indicate end-STSs from the clone under the symbol. Sizes of all clones are shown to scale. The chimeric part of YAC clone 856h2 (1,100 kb) is indicated by a dashed arrow. Interstitial deletions in YACs or PACs are indicated by thin lines in brackets. The extent of the deletion in YAC Y638E7 is not precisely known.

analysis of hybrid cell lines containing chromosome 12 as their only human chromosome. Map position in 12q24.1 for clones B087CO5, B55CO5, and P65I22 was confirmed by FISH analysis (data not shown).

The dense localization of STSs allowed the precise positioning of YACs in the region. Y884-h-11 was the only YAC clone that was positive for both *D12S1332* and *D12S1333*. However, this clone contained an interstitial deletion that encompassed approximately 200 kb. A small portion of this deletion, which was subsequently shown to contain parts of *SCA2*, was not covered by any of the other YAC clones, but was contained in several PAC clones (Fig. 1).

Genomic analysis of the *SCA2* repeat

We identified clones containing trinucleotide repeats by hybridizing *XbaI*/*NotI* digests of a minimal tiling path of clones with a $(CAG)_{10}$ oligonucleotide, as well as other trinucleotide permutations (data not shown). Two CAG positive bands of distinct sizes were identified in the contig. Sequence analysis of one of these after cloning into plasmid Pl65I22 contained an extended CAG repeat that was twice interrupted and had the following structure: $(CAG)_8CAA(CAG)_4CAA(CAG)_8$. The repeat was embedded in a long open reading frame (corresponding to basepairs 163 to 890 of the cDNA sequence shown in Fig. 2*a*) terminated by a putative splice site 3' of the repeat which was subsequently confirmed by comparison with the cDNA sequence.

To analyse genomic DNAs in normal individuals and SCA2 patients, the region containing the repeat was amplified using several primer pairs. The best results were obtained using primer pairs SCA2-A and -B. On agarose gels, a single band of approximately 130 bp was detected in normal individuals, whereas all patients with SCA2 from three independent pedigrees[11] showed one allele in the normal size range and a larger allele ranging from approximately 190 to 250 bp. Southern blot analysis confirmed that both alleles contained CAG repeats (data not shown).

To determine the exact sizes of amplified fragments, DNAs from SCA2 patients and normal individuals were amplified and PCR products separated by polyacrylamide gel electrophoresis (Fig. 3*a*). A common allele of 22 repeats and a less frequent allele of 23 repeats were seen. In 110 normal chromosomes from 55 Caucasian individuals of Northern, Southern and Eastern European origin, the allele frequencies were 0.92 for the smaller and 0.08 for the larger allele. Expanded alleles ranging from 36 to 52 repeats were observed in patients from three independent SCA2 pedigrees. Once expanded to the pathologic range, the SCA2 repeat was moderately unstable and further expansion by two to nine repeat units was seen during meiosis (Fig. 3*a*). There was great variability in the age of onset for a given repeat length, especially for disease alleles with 36–40 repeats. Due to heterogeneous variance in age of onset we used non-linear regression (Fig. 3*b*) — a negative exponential function was succesfully fitted (see Methods). The smallest expansion of 36 repeats was seen in two men with disease onset at ages 37 and 44. The longest expansion of 52 repeats was seen in a boy with disease onset at 9 years of age.

Sequence analysis of ten normal alleles revealed that the common normal allele with 22 repeats contained the two CAA interruptions that were also seen in plasmid

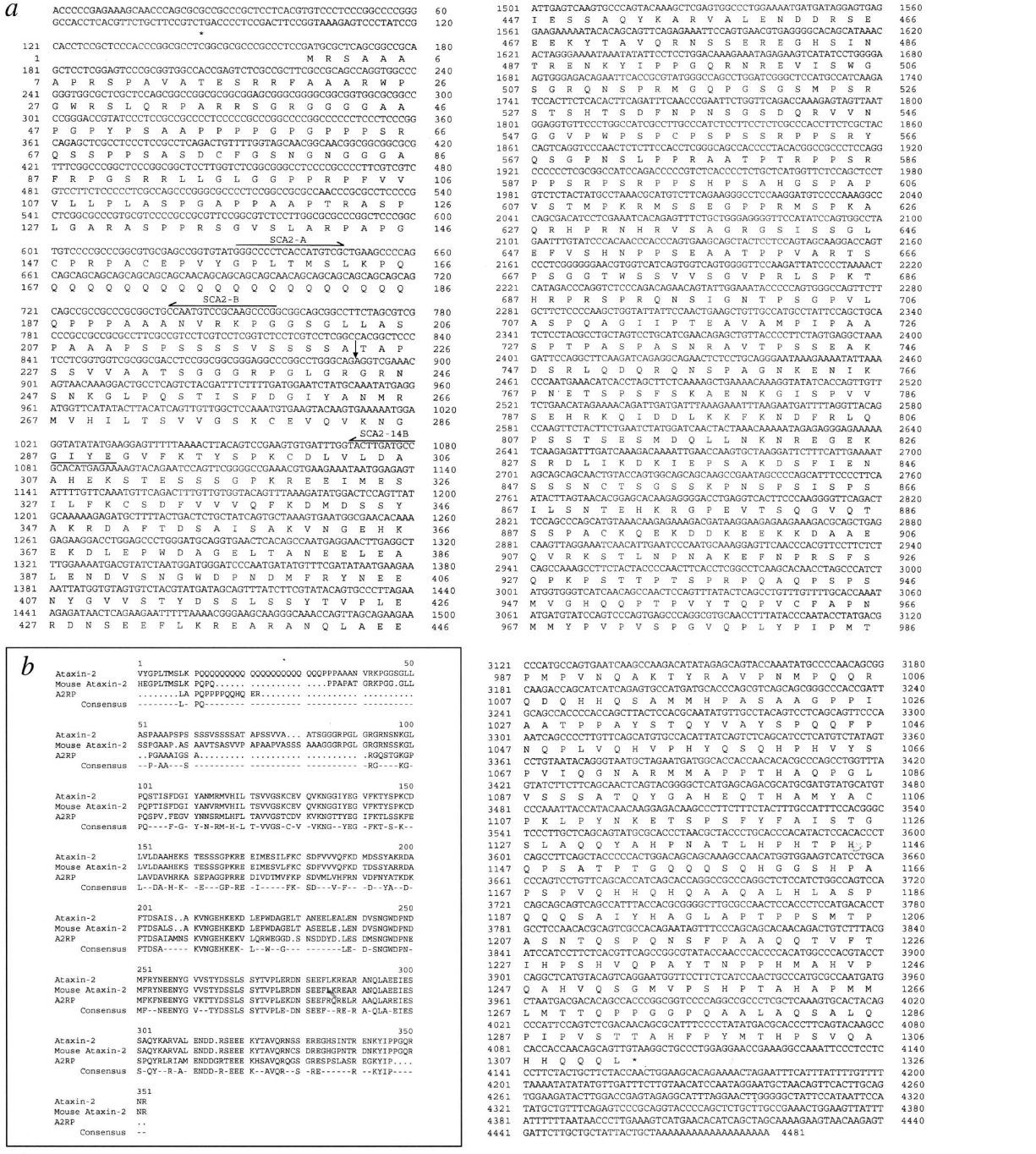

Fig. 2 *SCA2* cDNA and predicted amino acid sequence. *a*, Composite cDNA sequence obtained from assembly of partially overlapping cDNA clones. The predicted protein product is shown below the DNA sequence. In-frame stop codons preceding the first in-frame methionine and terminating the open reading frame are indicated by *. The locations of primers SCA2-A, SCA2-B, and SCA2-14B are indicated by horizontal arrows. The splice site between primers SCA2-B and SCA2-14B is indicated by a vertical arrow. *b*, Amino acid sequence alignment of ataxin-2, the ataxin-2 related protein, and the mouse SCA2 homologue in the region of strongest homology. Codon 1 corresponds to codon 155 in (*a*). The cDNA sequences for the three genes have been deposited in GenBank.

Pl65I22. The less frequent normal allele with 23 repeats had lost the 5' CAA interruption (Fig. 4), and contained an additional CAG repeat at the 5'-end of the repeat. In three expanded alleles that were isolated from SCA2 patients the CAG repeat lacked any interruptions.

Previous analysis of trinucleotide repeats predisposed to expansion had suggested that these regions are predicted to form hairpin structures[20]. We used an updated version of the DNA-FOLD Program[21] for secondary structure predictions which suggested the formation of several possible hairpin structures. Whereas the presence of two CAA interruptions results in a branched hairpin with a free energy of –21.4 kcal, the loss of interruption is predicted to result in a perfect hairpin with a free energy of –25.2 kcal (nucleotides 648–743, Fig. 2*a*).

To determine the frequency of mutation of *SCA2* in non-Portuguese patients we screened DNAs from 45 independent families with autosomal dominant SCAs. Expansion of the *SCA2* repeat was detected in six families. In this set of families, *SCA2* expansion was twice as

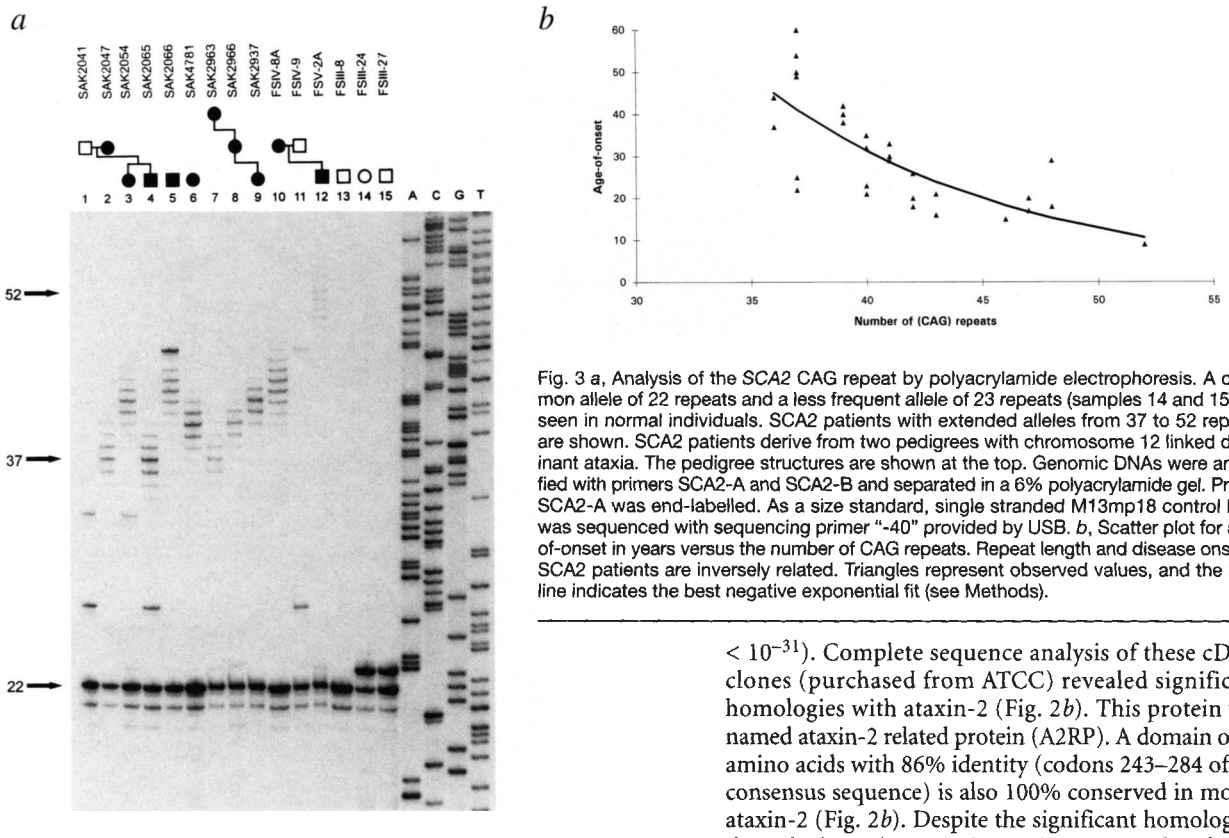

Fig. 3 *a*, Analysis of the *SCA2* CAG repeat by polyacrylamide electrophoresis. A common allele of 22 repeats and a less frequent allele of 23 repeats (samples 14 and 15) are seen in normal individuals. SCA2 patients with extended alleles from 37 to 52 repeats are shown. SCA2 patients derive from two pedigrees with chromosome 12 linked dominant ataxia. The pedigree structures are shown at the top. Genomic DNAs were amplified with primers SCA2-A and SCA2-B and separated in a 6% polyacrylamide gel. Primer SCA2-A was end-labelled. As a size standard, single stranded M13mp18 control DNA was sequenced with sequencing primer "-40" provided by USB. *b*, Scatter plot for age-of-onset in years versus the number of CAG repeats. Repeat length and disease onset in SCA2 patients are inversely related. Triangles represent observed values, and the bold line indicates the best negative exponential fit (see Methods).

common as expansion in the *SCA1* gene (D. Geschwind and S.-M.P., unpublished). In addition to individuals with a 'typical' SCA phenotype, expansion of the *SCA2* repeat was detected in a pedigree with a MJD phenotype and in one family with SCA and marked dementia.

cDNA clone isolation and *SCA2* gene expression

Using fragments generated by PCR amplification of Pl65I22 we screened a human adult frontal cortex library (Stratagene) and a human fetal brain library generated from a fetus with trisomy 21 (ref. 22). A total of 17 clones were obtained which appeared to belong to a total of four partially overlapping classes of clones. One set of clones extended from 265 bp 5' of the CAG repeat to the untranslated region and included a poly-A tail. The 5'-end of the *SCA2* cDNA was obtained by sequence analysis of cloned RT-PCR fragments.

The longest open reading frame consists of 3,936 bp and ends with a TAA termination codon (Fig. 2*a*). The stop codon is followed by 364 bp of 3' untranslated sequence. The CAG repeat is located in the 5' end of the coding region. The putative translation start site follows an in-frame stop codon located 78 bp upstream. The predicted molecular weight for the SCA2 translation product is 140.1 kD with the CAG trinucleotide repeat predicted to code for glutamine. In analogy to the *SCA1* gene product, we propose the name ataxin-2 for the *SCA2* gene product.

Comparison of this sequence against the GenBank database using the FASTA sequence alignment algorithm did not reveal significant similarities to genes of known function. However, significant similarities were detected with two partial cDNA transcripts in the TIGR database (THC148678, H03566, odds against chance similarity $< 10^{-31}$). Complete sequence analysis of these cDNA clones (purchased from ATCC) revealed significant homologies with ataxin-2 (Fig. 2*b*). This protein was named ataxin-2 related protein (A2RP). A domain of 42 amino acids with 86% identity (codons 243–284 of the consensus sequence) is also 100% conserved in mouse ataxin-2 (Fig. 2*b*). Despite the significant homologies, the polyglutamine tract in ataxin-2 was replaced with an interrupted polyproline tract in the related human protein and in the mouse homologue.

Using RT-PCR we determined that the *SCA2* CAG repeat was transcribed in lymphoblastoid cell lines (Fig. 5*a*). In cDNAs from SCA2 patients, transcription from both the normal and the expanded allele was detected using oligonucleotide primers that flank the repeat; the latter hybridize with sequence in different exons, avoiding amplification of genomic DNAs. Northern blot analysis revealed that *SCA2* was widely expressed. A strong signal corresponding to a 4.5-kb transcript was detected in RNAs isolated from brain, heart, placenta, liver, skeletal muscle, and pancreas (Fig. 5*b*). Little transcript was detected in lung or kidney. A much fainter transcript of 7.5 kb could also be seen in some tissues.

Discussion

We have identified the *SCA2* gene, a new gene containing an unstable CAG repeat. Unstable CAG repeats have been identified in SBMA, Huntington disease, DRPLA, SCA1 and SCA3. In contrast to the CTG repeat in myotonic dystrophy[23,24] and the CCG repeats associated with chromosomal fragile sites[25], the CAG repeats in these neurodegenerative disorders are in the coding region of the respective genes and result in extended polyglutamine tracts. SCA2 represents the sixth disease in which expansion of a CAG trinucleotide repeat causes disease, but there are several features of the *SCA2* repeat that appear to be unique.

The *SCA2* repeat is unusual in that only two alleles were seen in the normal population. A common allele with 22 repeats was found in 92% of chromosomes of people of European descent and a rare second allele in 8% of chromosomes. As we used only DNA samples from individuals of European descent, these allele fre-

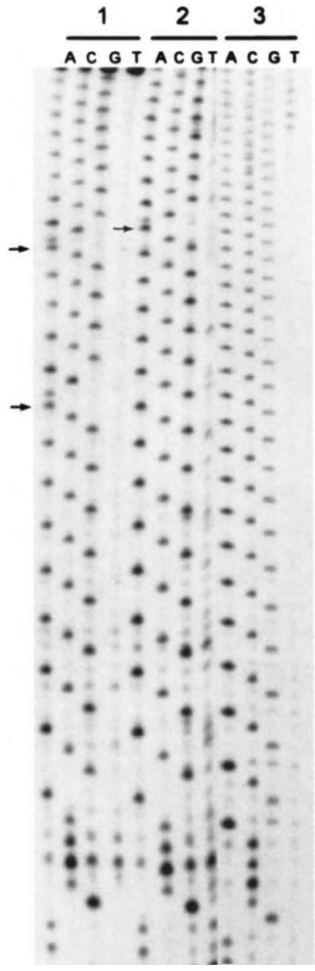

Fig. 4 DNA sequence analysis of the CAG repeat in normal individuals and SCA2 patients. The repeat in the common normal allele contains two CAA interruptions (arrows, lane 1) and one interruption in the less common normal allele (lane 2), but the repeat is perfect in a patient with SCA2 (lane 3). Genomic DNAs were amplified using primers SCA2-A and B, PCR products separated by agarose gel electrophoresis; bands were excised from the gel and sequenced using primer SCA2-A.

quencies should be analysed in other populations. In the other five CAG expansion diseases, the CAG repeats on normal chromosomes are highly polymorphic with a heterozygosity in the range of 0.80 and above. Repeat sizes on normal chromosomes range from a low of seven repeats in *DRPLA* to 40 repeats in *SCA3/MJD*[3,15]. It has been suggested that the extended normal alleles represent founder alleles which are predisposed to expansion[26,27]. Due to the rarity of *de novo* SCA patients we were not able to examine whether extended normal alleles close to the disease range exist that may provide a population of founder alleles. If they do exist, their frequency is much lower than the frequency of founder chromosomes in fragile X syndrome[28].

Expansion of the *SCA2* CAG repeat on disease chromosomes was relatively moderate and was in the range seen with expansions in the SBMA and HD genes. The lowest number of repeats causing SCA2 was 36 and the most common disease allele had 37 repeats. The longest normal and the shortest *SCA2* disease allele were separated by 13 repeats. Disease alleles showing 36 repeats have now clearly been established for HD[29], although normal elderly individuals with 36–40 repeats exist and the most common HD alleles have >40 repeats. Once expanded on disease chromosomes, the *SCA2* repeat may undergo moderate expansions (Fig. 3*a*). Future studies will need to examine whether the degree of expansion is dependent on parental origin of the disease chromosome.

The *SCA2* repeat is contained in a novel gene which is transcribed in several tissues including non-neuronal tissues (Fig. 5*a,b*). The gene product, ataxin-2, has a predicted molecular weight of 140 kD — in good agreement with the 150-kD protein observed using a monoclonal antibody to long polyglutamine tracts[17]. Despite the phenotypic overlap of SCA2 with SCA1 and SCA3, ataxin-2 has no homologies with other ataxins nor with huntingtin, or the DRPLA and SBMA proteins.

However, ataxin-2 showed significant homologies with another novel protein. A 42-amino acid domain was identified that was 86% identical between the two proteins. The potential functional importance of this domain was underscored by the fact that it was 100% conserved in the mouse SCA2 homologue (Fig. 2*b*). Interestingly, the polyglutamine tract was not conserved in either protein. As the pathogenesis of polyglutamine containing proteins is still poorly understood, the identification of functionally important domains adjacent to polyglutamine tracts may provide the potential for novel strategies to analyse the function of ataxin-2. A gain of function for the mutated ataxin-2 is supported by the fact that transcripts coding for mutated alleles are detected by RT-PCR (Fig. 5*a*).

Expansion of the *SCA2* repeat appears to be a common cause of a dominant SCA phenotype in non-Portuguese patients. When samples from 45 families with a dominant SCA phenotype were screened, samples from 6 independent pedigrees showed expansion of the *SCA2* repeat. It has been suggested that there are features specific to SCA2, but this assessment was limited

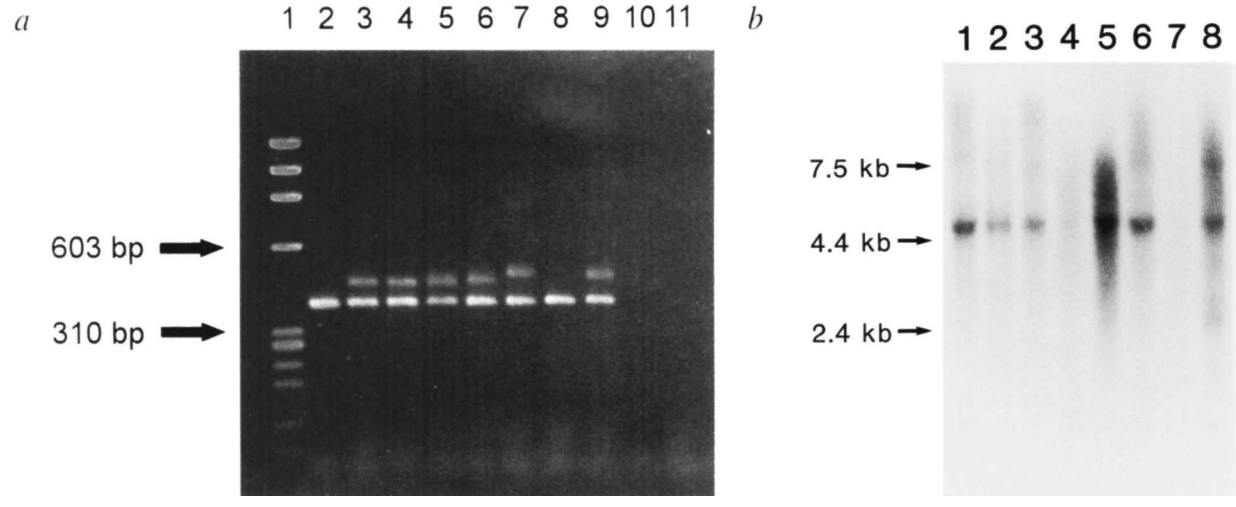

Fig. 5 Expression of the *SCA2* transcript as determined by reverse-transcribed PCR and northern blot analysis. *a*, Both normal and expanded *SCA2* alleles are expressed. RNAs isolated from normal individuals (lanes 2, 8) and SCA2 patients (lanes 3–7,9) were reverse transcribed into cDNA and amplified using primers SCA2-A and SCA2-14B. Primer SCA2-14B is located beyond the splice site indicated in Fig. 2*a*, and no amplification across the intervening intron is observed in genomic DNAs. Lane 1: F174/*Hae*III molecular weight ladder, lane 10: Genomic DNA from a SCA2 patient, lane 11: negative control without DNA. *b*, Northern blot of RNAs extracted from multiple human tissues probed with cDNA clones S1 and S2. A 4.5-kb transcript is detected in most tissues as well as a very faint 7.5-kb transcript. Lane 1, heart; lane 2, brain; lane 3, placenta; lane 4, lung; lane 5, liver; lane 6,skeletal muscle; lane 7, kidney; lane 8, pancreas.

to families large enough to be studied by linkage analysis[30]. A better assessment of the range of SCA2 phenotypes is now possible due to the ability to test small families and single cases. In our patient sample, most patients had a 'typical' SCA phenotype, but some patients had been classified as having an MJD phenotype and others showed a prominent dementia (D. Geschwind and S.-M.P., unpublished). Analysis of additional samples will be necessary to determine whether *SCA2* mutations can also cause HD or DRPLA-like phenotypes.

When performing direct testing for *SCA2* mutations, great caution has to be exercised when interpreting the presence of expanded *SCA2* alleles on polyacrylamide gels. A variable number of unrelated PCR fragments may be seen that are in the size range of expanded *SCA2* repeats. Although these bands lack the typical 'shadow' bands seen when di- or trinucleotide repeats are amplified, they may interfere with the interpretation in some samples. It is therefore recommended to confirm the presence of an expanded allele by Southern blotting and hybridization with a $(CAG)_{10}$ oligonucleotide.

Mechanisms of repeat expansion. Two types of expansion have been distinguished: class I expansion refers to increases of >10 repeats, while class II changes are expansions or contractions of <4 repeats. Large expansions are related to several features such as a critical repeat length, sequence of the repeat, and the ability of the repeat to form hairpin structures. Although direct evidence is still lacking, it has been suggested that distinct mechanisms for both types of expansions exist[31]. Class II changes may occur by simple DNA polymerase slippage or by a DNA hairpin mediated process, whereas class I changes occur only by the latter process.

For GC-rich repeats a critical free energy is only reached when at least 25 uninterrupted CAG repeats are found[20]. In the case of SCA2, the critical energy can likely be reached at a lower number of repeats due to the contribution of the GC-rich flanking region. Despite the considerable stability of the hairpin, the lack of CAA triplet interruptions in disease alleles confirms the importance of the sequence within the repeat similar to the lack of interruption of disease alleles in SCA1 and FMR1[28,32]. Presence of interruption in the *SCA2* repeat is predicted to result in a branched hairpin. Further studies are needed to determine if this may explain the lack of class II expansions. Analysis of the flanking sequence on normal and disease alleles may reveal whether other sequence variations can be identified that alter the pattern of major and minor hairpins. However, preliminary haplotype analysis of SCA2 families from France, Tunisia, Canada and the United States provided no support for a haplotype predisposed to expansion[9].

Mechanisms of disease. Although six human neurodegenerative diseases share extended polyglutamine stretches in the mutant proteins, the mechanisms underlying neurodegeneration by polyglutamine expansion are still controversial. A common mechanism is suggested by the phenotypic overlap between the polyglutamine diseases and by the identification of binding proteins that interact with polyglutamine peptides irrespective of the specific flanking amino acids. Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) is bound by a polyglutamine peptide, and both huntingtin and the *DRPLA* gene product bind to GAPDH *in vitro*[33]. Similarly, transfection of cells with a polyglutamine peptide induced apoptosis, whereas transfection with a full length *MJD* transcript did not[34]. On the other hand, a binding protein of huntingtin has been identified which appears to interact specifically with huntingtin and not with atrophin-1, another polyglutamine-containing protein[35]. Although huntingtin is widely expressed, the huntingtin-associated protein shows a pattern of expression restricted to cell types involved in HD and may explain the cell-type specific degeneration seen in HD.

An argument against a simple action purely based on length of the polyglutamine tract now comes from study of the ataxin-2 polyglutamine tract. It is noteworthy that the smallest number of glutamines seen in SCA2 patients (36 and 37 glutamines) is in the range of the longest polyglutamine tracts in the *SCA3/MJD* gene in normal individuals. The most common disease causing polyglutamine tract in ataxin-2 is actually three glutamines shorter than the longest tract seen in the *MJD* gene product in normal individuals. This argues against an interaction of polyglutamines with one target protein or glutamine toxicity simply based on length of the polyglutamine stretch and may point to a role played by specific flanking amino acids. It is likely that several proteins may serve as targets for the binding of polyglutamine proteins, and that overlapping subsets of binding proteins explain disease specific phenotypes. Intra- as well as interfamilial variability might be explained by polymorphism in these proteins or differential regulation of their expression. The identification of the *SCA2* gene will now permit testing of identified binding proteins for their ability to bind to ataxin-2, but may also lead to the identification of novel binding proteins.

In summary, the finding of CAG expansion in a novel gene underscores the importance of this mutational mechanism for neurodegeneration. Although class I expansions of the *SCA2* repeat are common, the lack of frequent class II expansions may be related to the unusual structure of the sequences flanking the *SCA2* repeat. Further analysis of the *SCA2* repeat in humans and its introduction into the animal germline may provide novel insights into stability of trinucleotide repeats.

Methods

PAC and BAC library screens. A 3x human PAC library was arrayed in 384 well dishes[18]. As a starting point to assemble a contig we screened PCR pools of the PAC library with the marker for *D12S1332*. Subsequent 'walking steps' were undertaken by hybridizing PCR-generated PAC-end STS fragments to gridded membranes of the 3x PAC library and a 1x total human genome bacterial artificial chromosome (BAC) library (Research Genetics). In a similar fashion, a second contig was established starting with *D12S1333*. All STSs were mapped back to human chromosome 12 by PCR analysis of a human/Chinese hamster somatic hybrid cell line, HHW582, which contains chromosome 12 as the only human chromosome, and by analysis of an extract from a chromosome 12 specific lambda library, LL12NS01 (both from Coriell Cell Repositories).

YAC, PAC and BAC DNA preparation. Yeast artificial chromosome (YAC) clones were obtained from the CEPH mega-YAC library and grown under standard conditions[36]. PAC and BAC clones were grown overnight in LB media containing 12.5 µg/ml kanamycin for PACs and 12.5 µg/ml chloramphenicol for BACs. DNAs were prepared by the alkaline lysis method. PAC

and BAC DNAs were digested with *Not*I and subjected to pulsed-field gel electrophoresis. Sizes were determined relative to λ concatamers.

Analysis by pulsed-field gel electrophoresis. Agarose plugs of yeast cells containing total YAC DNA were prepared[37] and subjected to pulsed-field gel electrophoresis in 1% SeaKem agarose in 0.5x TBE using the CHEF DRII Mapper (Bio-Rad). Gels were blotted onto Magna NT Nylon membranes using alkaline blotting, UV cross linked and baked at 80 °C for 2 h. Membranes were hybridized with total human DNA, washed according to standard procedures, and exposed to Kodak XAR5 film. PAC and BAC clones were sized after digestion with *Not*I. The sizes of individual clones were determined by comparison to their relative positions with molecular weight standards.

Analysis by fluorescence *in situ* hybridization (FISH). PAC or BAC clones were biotinylated by nicktranslation in the presence of biotin-14-dATP using the BioNick Labelling Kit (Gibco-BRL). FISH was performed as described[38]. The colour images were captured by using a Cooled-CCD camera (Pjotometrics) and BDS image analysis software (Oncor Imaging, Inc.).

Sequencing of PAC endclones. PAC clones were inoculated into 500 ml of LB/kanamycin and grown overnight. DNAs were isolated using QIAGEN columns according to the vendor's protocol with one additional phenol/chloroform/isoamylalcohol extraction followed by an additional chloroform/isoamylalcohol extraction. Clones were sequenced using the cycle sequencing kit (Gibco-BRL) with standard T7 and SP6 primers.

Hybridization of $(CAG)_{10}$ oligonucleotides. Oligonucleotides (80 ng) were 5' end-labelled and hybridized overnight at 42 °C in buffer containing 1 M NaCl, 0.05 M Tris-HCl pH 7, 5.5 mM EDTA, 0.1 % SDS, 1x Denhardt's solution and 200 µg/ml denatured salmon sperm DNA. Filters were washed twice with 2× SSC, 0.1% SDS at 55 °C and exposed to Kodak X-ray film for 24 h, and subsequently washed at 65 °C, followed by additional exposure to X-ray film.

Cloning and sequencing of the *SCA2* CAG-repeat. PAC clone 65I22 was digested with *Sau*3AI and cloned into the pBluescript SK(+) phagemid (Stratagene). After transfection into DH5α, bacterial colonies were screened for poly-CAG containing inserts using the methods described above. Positive clones were sequenced using the CircumVent Thermal Cycle Dideoxy DNA sequencing kit (New England Biolabs) with end-labelled T3 and T7 primers.

PCR conditions. Eighty ng each of primers SCA2-A (5'–GGGCCCCTCACCATGTCG–3') and SCA2-B (5'–CGGGCTTGCGGACATTGG–3') were added to 20 ng of human DNA with standard PCR buffer and nucleotide concentrations. After an initial denaturation at 95 °C for 5 min, 35 cycles were repeated with denaturation at 96 °C for 1.5 min, an annealing temperature of 63 °C for 30 s, extension at 72 °C for 1.5 min and a final extension of 5 min at 72 °C.

***SCA2* repeat analysis by polyacrylamide electrophoresis.** The SCA2-A oligonucleotide primer was end-labelled at the 5' end with [γ–^{32}P]ATP and utilized in PCR reactions as described above except that the amount of primer reduced to SCA2-A was 7 ng. PCR products were separated by electrophoresis through 6% polyacrylamide DNA sequencing gels.

Isolation of cDNA clones. ^{32}P-labelled probes were generated by PCR amplification of plasmid Pl65I22 using the following primer pair: 65A3: 5'–CCGCGGCTGCCAATGTCC–3', 65B5: 5'–GTAACCGTTCGGCGCCCG–3'. A second probe was generated using primers 65A6: 5'GGCTCCCGGCGGCTCCTT–3'; 65B6: 5'–TGCTGCTGCTGCTGGGGCTTCAG–3'. These were used to screen the Stratagene adult human frontal cortex Lamba Zap II cDNA library and a trisomy 21 fetal brain cDNA library[22]. Two clones (designated S1 and S2) were identified in the first library and 15 clones (designated F1.1–F1.7, and F2.1–F2.8) in the second. Phages were plated to an average density of 1×10^5 per 150 mm plate. Plaque lifts of 20 plates (2×10^6 phages) were made using duplicated nylon membranes (Duralose-UV, Stratagene). Hybridization and excision were performed according to the manufacturer's protocol. Hybridized membranes were washed to a final stringency of 0.2x SSC, 0.1x SDS at 65 °C. The filters were exposed overnight onto X-ray film. Excised phagemids were grown overnight in 5ml LB medium containing 50 ug/ml of ampicillin. To obtain cDNA sequence for the 5' end, placental poly-T selected placental mRNAs (Clontech) were transcribed with MMLV reverse transcriptase and amplified with the following primers: SCA2-A30: 5'–CCGCCCGCTCCTCACGTGT–3'; SCA2-A31: 5'–ACCCCCGAGAAAGCAACC–3'; SCA2-B30: 5'–CCGTTGCCGTTGCTACCA–3'. The sequences for primers SCA2-A30 and A31 were obtained from genomic sequence, and are located 5' to the stop codon preceding the putative initiator methionine. The sequence for SCA2-B30 was obtained from the 5' end of cDNA clones F1.1 and F1.2. The amplicons obtained by RT-PCR were directly sequenced. To identify mouse *SCA2* cDNA clones, we screened the Stratagene Lambda ZAP newborn mouse brain cDNA library with a human *SCA2* cDNA clone. Six clones were identified and sequenced.

Northern blot and RT-PCR analysis. Multiple tissue northern blots were purchased from Clontech. RNAs were isolated from lymphoblastoid cell lines established from patients and unrelated spouses in the FS pedigree[5] and reverse transcribed as previously described[39]. For amplification, primers located in two exons (SCA-A and SCA-14B, see also Fig. 2*a*) were chosen so that genomic DNA was not amplified. The sequence for SCA-14B was: 5'–TTCTCATGTGCGGCATCAAG–3'.

DNA secondary structure analysis. We used an updated version of DNA-FOLD that incorporated improved nearest neighbor parameters for calculations[21]. Therefore free energies cannot be directly compared with energies calculated for other expanding CAG repeats[20].

Sequence alignment. Amino acid sequences were aligned using the PILEUP routine from GCG (Genetic Computer Group).

Regression analysis. The data were fit using the Statistical Analysis Software (SAS) package version 3.10 using the Secant Method[40]. The regression equation was y = A*exp (–ax), where y gives the age of onset and x the number of CAG repeats. The conversion criteria were met with the mean square error of 76.598. The values of parameters are as follows: A=1171.583, a=0.091.

GenBank accession numbers. *SCA2* cDNA: U70323; human ataxin-2 related protein cDNA: U70671; mouse *SCA2* cDNA: U70670.

Acknowledgements

We thank T. Hudson for the gift of YACs Y884h11, Y919a2, Y792b4, Y731b7, and Y856h2, and K. Yamakawa for the fetal human brain cDNA library. This work was supported by the Carmen and Louis Warschaw Endowment Fund for Neurology, the Ruth and Lawrence Harvey Endowment Fund for the Neurosciences, the National Institutes of Health (all to S.-M. P.), grants from DFG and BMBF (to G.A.), grants from the Department of Energy and the Brawerman Fund for Molecular Genetics (to J.R.K.) and the joint Program Fonds de la Recherche en Santé du Québec and Association Canadienne de l'Ataxie de Friedreich (to G.A.R.).

Received 7 August; accepted 23 September 1996.

1. Gudmundsson, K. The prevalence and occurrence of some rare neurological diseases in Iceland. *Acta. Neurol. Scan.* **45**, 114–118 (1969).
2. Orr, H.T. *et al.* Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. *Nature Genet.* **4**, 221–226 (1993).
3. Kawaguchi, Y. *et al.* CAG expansions in a novel gene for Machado–Joseph disease at chromosome 14q32.1. *Nature Genet.* **8**, 221–228 (1994).
4. Gispert, S. *et al.* Chromosomal assignment of the second locus for autosomal dominant cerebellar ataxia (SCA) to chromosome 12q23–24.1. *Nature Genet.* **4**, 295–299 (1993).
5. Pulst S.M., Nechiporuk A. & Starkman, S. Anticipation in spinocerebellar ataxia type 2. *Nature Genet.* **5**, 8–10 (1993).
6. Flanigan, K. *et al.* Autosomal dominant spinocerebellar ataxia with sensory axonal neuropathy (SCA4): Clinical description and genetic localization to chromosome 16q22.1. *Am. J. Hum. Genet.* **59**,392–399 (1996).
7. Ranum L.P.W., Schut, L.J., Lundgren, J.K. & Orr, H.T. Spinocerebellar ataxia type 5 in a family descended from the grandparents of president Lincoln maps to chromosome. *Nature Genet.* **8**, 280–284 (1994).
8. Gouw L.G. *et al.* Retinal degeneration characterizes a spinocerebellar ataxia mapping to chromosome 3p. *Nature Genet.* **10**, 89–93 (1995).
9. Gispert, S. *et al.* Localization of the candidate gene D–Amino acid oxidase outside the refined 1–cM region of spinocerebellar ataxia 2. *Am. J. Hum. Genet.* **57**, 972–975 (1995).
10. Krauter, K. *et al.* A second–generation YAC contig map of human chromosome 12. *Nature* **377**, 321– 323 (1995).
11. Nechiporuk, A. *et al.* Genetic mapping of the spinocerebellar ataxia type 2 gene on human chromosome 12. *Neurology* **46**, 1731–1735 (1996).
12. Durr, A. *et al.* Dominant cerebellar ataxia type I linked to chromosome 12q (SCA2: spinocerebellar ataxia type 2). *Clin. Neurosci.* **3**, 12–16 (1995).
13. 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).
14. La Spada, A.R., Wilson, E.M., Lubahn, D.B., Harding, A.E., & Fischbeck K.H. Androgen receptor gene mutations in X–linked spinal and bulbar muscular atrophy. *Nature* **352**, 77–79 (1991).
15. Koide, R. *et al.* Unstable expansion of CAG repeat in hereditary dentatorubral–pallidoluysian atrophy (DRPLA). *Nature Genet.* **6**, 9–13 (1994).
16. Nagafuchi S *et al.* Dentatorubral and pallidoluysian atrophy expansion of an unstable CAG trinucleotide on chromosome 12p. *Nature Genet.* **6**, 14–18 (1994).
17. Trottier, Y. *et al.* Polyglutamine expansion as a pathological epitope in Huntington's disease and four dominant cerebellar ataxias. *Nature* **378**, 403–406 (1995).
18. Ioannou, P.A. *et al.* A new bacteriophage P1–derived vector for the propagation of large human DNA fragments. *Nature Genet.* **6**, 84–89 (1994).
19. Nechiporuk, T. *et al.* Identification of three new microsatellite markers in the spiocerebellar ataxia type 2 (SCA2) region and 1.2 Mb physical map. *Hum. Genet.* **97**, 462–467 (1996).
20. Gacy, A.M., Goellner G., Juranic N., Macura S. & McMurray, C.T. Trinucleotide repeats that expand in human disease form hairpin structures *in vitro*. *Cell* **81**, 533–540 (1995).
21. SantaLucia, J. Jr., Allawi, H.T. & Seneviratne, P.A. Improved nearest–neighbor parameters for predicting DNA duplex stability. *Biochemistry* **35**, 3555–3562 (1996).
22. Yamakawa, K. *et al.* Isolation and characterization of a candidate gene for progressive myoclonus epilepsy on 21q22.3. *Hum. Mol. Genet.* **4**, 709–716 (1995).
23. Brook J.D. *et al.* Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3' end of a transcript encoding a protein kinase family member. *Cell* **68**, 799–808 (1992).
24. Fu, Y.H. *et al.* An unstable triplet repeat in a gene related to myotonic muscular dystrophy. *Science* **255**, 1256–1258 (1992).
25. Nelson D. The fragile X syndromes. *Cell Biol.* **6**, 5–11 (1995).
26. Goldberg, Y.P. *et al.* Molecular analysis of new mutations for Huyntington's disease: intermediate alleles and sex of origin effects. *Nature Genetics* **5**, 174–179 (1993)
27. Myers, R.H. *et al. De Novo* expansion of a (CAG) repeat in sporadic Huntington's disease. *Nature Genet.* **5**, 168–173 (1993).
28. Kunst, C.B. & Warren S.T. Cryptic and polar variation of the fragile X repeat could result in predisposing normal alleles. *Cell* **77**, 853–861 (1994).
29. Rubinsztein, D.C. *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 to 39 repeats. *Am. J. Hum. Genet.* **59**, 16–22 (1996).
30. Filla, A. *et al.* Has spinocerebellar ataxia type 2 a distinct phenotype? Genetic and clinical study of an Italian family. *Neurology* **45**, 793–796 (1995).
31. McMurray, C.T. Mechanisms of DNA expansion. *Chromosoma* 104, 2–13 (1995).
32. Chung, M.Y., Ranum, L., Duvick, L., Servadio, A., Zoghbi, H. & Orr, H.T. Evidence for a mechanism predisposing to intergenerational CAG repeat instability in spinocerebellar ataxia type 1. *Nature Genet.* **5**, 252–258 (1993).
33. Burke, J.R. *et al.* Huntington and DRPLA proteins selectively interact with the enzyme GAPDH. *Nature Med.* **2**, 347–350 (1996).
34. Ikeda, H. *et al.* Expanded polyglutamine in the Machado–Joseph disease protein induces cell death *in vitro* and *in vivo*. *Nature Genet.* **13**, 196–202 (1996).
35. Li, X.–J. *et al.* A huntingtin–associated protein enriched in brain with implications for pathology. *Nature* **378**:398–402 (1995).
36. Cohen, D., Chumakov, I. & Weissenbach, J. A first–generation physical map of the human genome. *Nature* **366**, 698–701 (1993).
37. Larin, Z. & Lehrach, H. Yeast artificial chromosomes: an alternative approach to the molecular analysis of mouse developmental mutations. *Genet. Res.* **56**, 203–208 (1990).
38. Korenberg, J.R. & Chen, X.N. Human cDNA mapping using a high resolution R–banding technique and fluorescence *in situ*–hybridization. *Cytogenet. Cell Genet.* **69**, 196–200 (1995).
39. Huynh, D., Nechiporuk, T. & Pulst, S.–M. Alternative transcripts in the mouse neurofibromatosis type 2 (*NF2*) gene are conserved and code for schwannomins with distinct C–terminal domains. *Hum. Mol. Genet.* **3**, 1075–1079 (1994a).
40. Ralston, M.L. & Jennrich, R.I. DUD, a derivative free algorithm for non–linear least squares. *Technometrics* **20**, 7–14 (1978).