Human Molecular Genetics, 2001, Vol. 10, No. 21 2437–2446

CAG repeat instability at SCA2 locus: anchoring CAA interruptions and linked single nucleotide polymorphisms

Shweta Choudhry[1], Mitali Mukerji[1], Achal K. Srivastava[2], Satish Jain[2] and Samir K. Brahmachari[1,3,*]

[1]Functional Genomics Unit, Centre for Biochemical Technology (CSIR), Mall Road, Delhi, India, [2]Department of Neurology, Neurosciences Center, All India Institute of Medical Sciences, New Delhi, India and [3]Molecular Biophysics Unit, Indian Institute of Science, Bangalore, India

Received June 20, 2001; Revised and Accepted August 7, 2001 DDBJ/EMBL/GenBank accession nos AF330028–AF330033

Spinocerebellar ataxia 2 (SCA2) is an autosomal dominant neurodegenerative disorder that results from the expansion of a cryptic CAG repeat within the exon 1 of the SCA2 gene. The CAG repeat in normal individuals varies in length from 14 to 31 repeats and is frequently interrupted by one or more CAA triplets, whereas the expanded alleles contain a pure uninterrupted stretch of 34 to 59 CAG repeats. We have previously reported the presence of a limited pool of 'ancestral' or 'at risk' haplotypes for the expanded SCA2 alleles in the Indian population. We now report the identification of two novel single nucleotide polymorphisms (SNPs) in exon 1 of the SCA2 gene and their characterization in 215 normal and 64 expanded chromosomes. The two biallelic SNPs distinguished two haplotypes, GT and CC, each of which formed a predominant haplotype associated with normal and expanded SCA2 alleles. All the expanded alleles segregated with CC haplotype, which otherwise was associated with only 29.3% of the normal chromosomes. CAA interspersion analysis revealed that majority of the normal alleles with CC haplotype were either pure or lacked the most proximal 5′ CAA interruption. The repeat length variation at SCA2 locus also appeared to be polar with changes occurring mostly at the 5′ end of the repeat. Our results demonstrate that CAA interruptions play an important role in conferring stability to SCA2 repeat and their absence predisposes alleles towards instability and pathological expansion. Our study also provides new haplotypes associated with SCA2 that should prove useful in further understanding the mutational history and mechanism of repeat instability at the SCA2 locus.

INTRODUCTION

Spinocerebellar ataxias (SCAs) are a clinically heterogeneous group of autosomal dominant neurodegenerative disorders characterized by progressive deterioration in balance and coordination. The symptoms occur due to progressive neuronal loss primarily in the cerebellum, but also in other parts of the central nervous system. Eight disease loci have been identified to date as causing this phenotype: spinocerebellar ataxia 1 (SCA1), SCA2, Machado–Joseph disease (MJD)/SCA3, SCA6, SCA7, SCA8, SCA10 and SCA12 (1–10). The causative mutation associated with all these disease types is abnormal expansion of a trinucleotide repeat motif in their corresponding gene, except for SCA10, which is due to a pentanucleotide (ATTCT) repeat expansion. These repeats at different SCA loci are highly polymorphic in normal individuals, but once the repeat number crosses a particular threshold, specific to an individual locus, instability is observed leading to the manifestation of the disease (1–10).

SCA2, which was initially described in a Cuban population (11), has now been reported worldwide and is a quite frequent form of autosomal dominant cerebellar ataxia in various populations (12–16). The molecular basis for the disease is an expansion of a CAG repeat tract in exon 1 of the SCA2 gene located on chromosome 12q24.1 (2–4). In normal individuals the CAG repeat is not only polymorphic in length, ranging from 14 to 31 repeats, but also cryptic in nature, having one or more interruptions of CAA triplets (2–4,12–16). In contrast, the expanded SCA2 alleles contain a pure, continuous stretch of 34 to 59 CAG repeats (2–4,12–16). Despite the presence of CAA interruptions and variability in their number, most studies of SCA2 CAG repeat in human populations have focused on repeat length alone (12–16). Sequence interruptions within human di- and trinucleotide repeat tracts have been shown to play an important role in conferring increased genetic stability to the repeat tracts (17–20). The absence of these interruptions at least in fragile X syndrome and SCA1 predisposes alleles to expansion and eventually to disease status (21–24). It has been postulated that interruptions provide genetic stability to the repeat tracts by inhibiting strand slippage, such that upon transmission the interrupted tracts are less likely to expand (20,25). The observation that the majority of the normal SCA2 alleles possess an interrupted repeat configuration, whereas the disease alleles consist of pure CAG repeat tracts, suggests that it is important to determine both the length and the internal

*To whom correspondence should be addressed at: Functional Genomics Unit, Centre for Biochemical Technology (CSIR), Delhi University Campus, Mall Road, Delhi 110 007, India. Tel: +91 11 7416489; Fax: +91 11 7667471; Email: skb@cbt.res.in

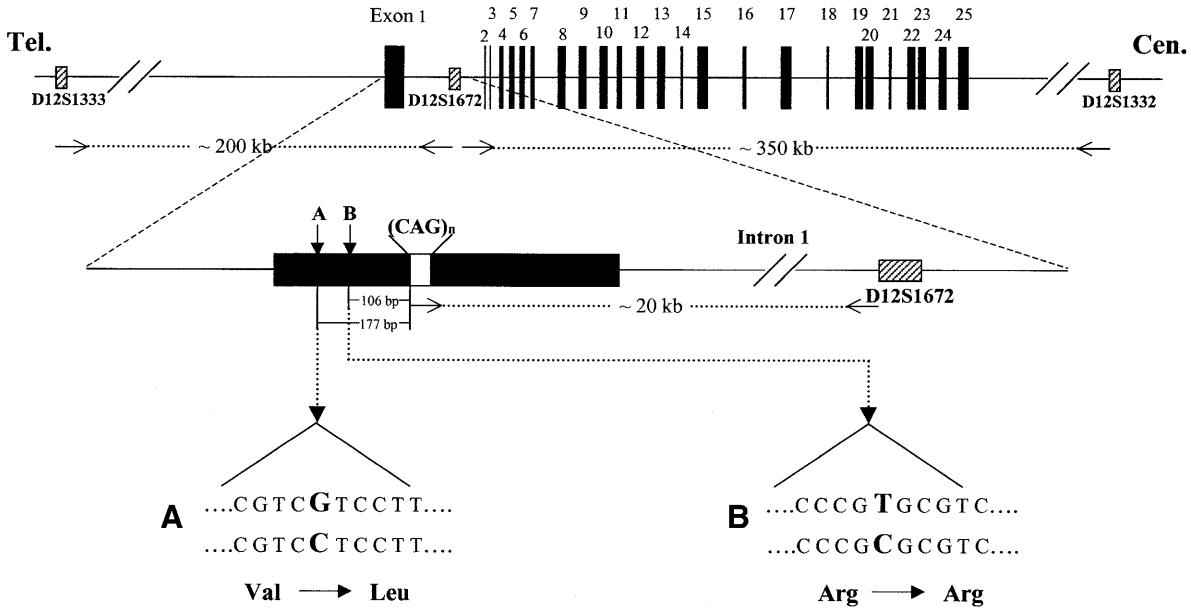


Figure 1. Schematic presentation of microsatellite markers and SNPs used for haplotype analysis. The top line depicts the positions of 25 exons of the SCA2 gene (34) and three microsatellite markers, D12S1333, D12S1672 and D12S1332. The second line shows the relative locations of two novel SNPs (**A** and **B**) and CAG repeat tract in exon 1 of the SCA2 gene. The first SNP (A) is located 177 bp upstream of the CAG repeat and is scored as either a G or C. The second SNP (B) is present 106 bp upstream of the SCA2 repeat tract and contains either a T or a C base. Both the polymorphisms are also shown in sequence context below the gene.

structure of SCA2 repeat alleles to evaluate their tendency for instability.

SCA2 has been shown to be the most frequent cause of dominantly inherited ataxia in the Indian population (16,26,27). Haplotype studies carried out using microsatellite markers flanking the SCA2 repeat have provided the support for a limited pool of 'ancestral' or 'at risk' haplotypes from which SCA2 disease chromosomes are derived in our population (16,28). In this paper, we report the identification and characterization of two novel single nucleotide polymorphisms (SNPs) in exon 1 of the human SCA2 gene. Using these polymorphic markers, we have extended our previous study of SCA2 in the Indian population (16) and investigated the role of various genetic factors such as repeat length, CAA interruption pattern and haplotype background in predisposing repeats towards instability and pathological expansion at the SCA2 locus. Based on our results we demonstrate that the instability of SCA2 CAG repeat alleles is likely to result from the absence of anchoring CAA interruptions.

RESULTS

Polymorphic markers around the SCA2 CAG repeat

In order to amplify the SCA2 CAG repeat containing region by PCR we had designed and tested three primer sets on normal and SCA2 positive samples. Sequence analysis of the PCR products obtained using one of these primer sets, SCA2-FP3 and SCA2-RP3 (Materials and Methods), revealed two novel intragenic, biallelic SNPs in exon 1 of the human SCA2 gene. The two SNPs were at nucleotide positions 481 and 552 in the human SCA2 mRNA sequence (GenBank accession no. U70323; www.ncbi.nlm.nih.gov/Genbank/). The first polymorphic site, as shown in Figure 1A, had either a G or a C base and was 177 bp upstream of the polymorphic CAG repeat stretch (SNP database reference SNP Id no. 695871; www.ncbi.nlm.nih.gov/SNP/). The second polymorphic site (Fig. 1B) was situated 106 bp upstream of the CAG repeat tract and contained either a T or a C base (SNP database reference SNP Id no. 695872). While the first substitution changes the amino acid sequence from valine to leucine, the second substitution is neutral.

SNPs and SCA2 CAG repeat length

We analyzed the two SNPs and the SCA2 CAG repeat length polymorphism in 215 normal and 64 expanded chromosomes. The phases of the two SNP alleles were determined based on their segregation pattern (Materials and Methods). Although four haplotypes are possible with two biallelic polymorphic systems, only two were observed: GT or CC haplotype. No GC or CT haplotype was detected in any sample, suggesting that either these haplotypes are very rare or G, T and C, C are exclusively linked to each other. The frequency of each SNP haplotype and its association with CAG repeat number is summarized in Table 1. In 215 normal chromosomes analyzed, the GT and the CC haplotypes were represented in 70.7 and 29.3%, respectively. Since only ~8% of the normal chromosomes at SCA2 locus contain CAG repeat number other than 22, we also compared the frequency distribution of the SNP haplotypes in chromosomes with 22 repeats and chromosomes having other than 22 repeats. Both these CAG repeat categories were significantly different (Fisher's exact test,

Table 1. Frequency distribution of SNP haplotypes in normal and expanded SCA2 chromosomes

CAG repeat length	*n*	Percentage GT haplotype (n)	Percentage CC haplotype (n)
Normal (18–30 repeats)	215	70.7% (152)	29.3% (63)
Repeats (= 22)	175	70.9% (124)	29.1% (51)
Repeats (other than 22)	31	22.6% (7)	77.4% (24)
Expanded (35–48 repeats)	64	0.0% (0)	100% (64)

$P = 0.0000$) for the two SNP haplotypes, the chromosomes with CC haplotype being much more frequent in the group with repeat other than 22 (77.4%) than with 22 repeats (29.1%) (Table 1). Further studies revealed a highly significant difference ($\chi^2 = 99.40$, $P < 0.0001$) in the distribution of the two SNP haplotypes between the normal and the expanded SCA2 chromosomes (Table 1). All the expanded chromosomes segregated with CC allele showing that disease chromosomes are in complete association with CC haplotype.

Association of SNP haplotype and flanking microsatellite markers

A common disease haplotype has been observed for Indian SCA2 pedigrees using flanking microsatellite markers D12S1333, D12S1672 and D12S1332 (locations shown in Fig. 1) (16,28). In order to examine the association between the microsatellite markers and the SNP alleles, the haplotype analysis was carried out in 17 SCA2 pedigrees and 200 normal control chromosomes. The disease haplotype segregating in each of the 17 pedigrees is shown in Table 2. A marked association was observed between the disease locus and the SNP haplotype, with all the disease chromosomes segregating with CC haplotype. Strong association was also observed between SCA2 and allele 1 (225 bp) at D12S1333 locus, alleles 3 (281 bp) and 4 (283 bp) at D12S1672 locus and allele 8 (208 bp) at D12S1332 locus (Table 3). The haplotype 1-3-8 for markers D12S1333, D12S1672 and D12S1332 was associated with SCA2 expansion in six and the truncated haplotype 1-3 for markers D12S1333 and D12S1672 in 11 out of 17 SCA2 families studied (Table 3).

A significant difference in the frequency distribution of the microsatellite alleles associated with SCA2 expansion mutation was observed for the normal chromosomes when grouped by SNP haplotype (Table 4). For example, allele 1 of D12S1333 marker was significantly (Fisher's exact test, $P = 0.0033$) over-represented in normal chromosomes with CC haplotype (25%) than with GT haplotype (7%). Similarly allele 3 of D12S1672 marker was present in much higher frequency in CC chromosomes (44%) compared with GT (3%). Despite a significant association of allele 8 at D12S1332 locus with disease chromosomes, no apparent difference was observed between the frequency distribution of this allele in GT (8%) and CC (7%) chromosomes. The truncated haplotype 1-3, which accounted for ~65% of the SCA2 families, was associated with only two (2%) control chromosomes, both having CC haplotype (Table 4).

Table 2. Haplotype analysis of SCA2 pedigrees

Pedigree	SNP haplotype	*D12S1333*	*D12S1672*	*D12S1332*
AT006	CC	1	3	8
AT009	CC	1	3	8
AT010	CC	1	3	6
AT024	CC	1	3	5
AT027	CC	1	3	8
AT031	CC	1	3	5
AT038	CC	1	3	8
AT048	CC	1	3	8
AT049	CC	1	3	8
AT056	CC	1	3	6
AT092	CC	1	3	7
AT017	CC	1	4	5
AT029	CC	1	4	5
AT052	CC	1	4	8
AT077	CC	1	4	5
AT041	CC	1	5	5
AT064	CC	3	4	5

Seventeen unrelated SCA2 pedigrees of Indian origin were typed using SNPs and microsatellite markers. Some parts of this table are derived from previously reported studies (16).
D12S1333: allele 1 = 225 bp, allele 3 = 229 bp; D12S1672: allele 3 = 281 bp, allele 4 = 283 bp, allele 5 = 285 bp; D12S1332: allele 5 = 202 bp, allele 6 = 204 bp, allele 7 = 206, allele 8 = 208 bp.

SNP haplotype and CAA interruption pattern

In order to assess the allelic diversity at SCA2 locus and to determine the effect of CAA interruptions on repeat stability, 215 normal and 64 expanded chromosomes were analyzed for the interruption pattern of CAA triplets within the SCA2 CAG repeat tract. Considerable variation was observed in the cryptic nature of the SCA2 repeat in normal chromosomes (Table 5). Two interruption patterns $(CAG)_8CAA(CAG)_4CAA(CAG)_8$ or 8+4+8 and $(CAG)_{13}CAA(CAG)_8$ or 13+8 (Materials and Methods for nomenclature) were observed most frequently, accounting for 70.7% (152/215) and 20.5% (44/215) alleles, respectively. Among 215 control chromosomes, 157 (73.0%) had two CAAs, 53 (24.7%) had one CAA, four (1.8%) were devoid of interruptions and one (0.5%) had three CAA interruptions. When both the repeat length and the CAA interruption pattern were considered, the heterozygosity at SCA2 locus was much higher (51%) for our sample set compared with 13% based on repeat length alone. While 98% (211/215) of the normal chromosomes had an interrupted repeat configuration, sequence analysis of 64 expanded SCA2 alleles revealed a homogeneous, uninterrupted CAG repeat stretch.

A marked split was observed in the number and the pattern of CAA interruptions between the chromosomes with GT and CC haplotype (Fig. 2). Ninety-eight percent (149/152) of the chromosomes with GT haplotype had two or more CAA interruptions while 86% (54/63) of the CC chromosomes had either

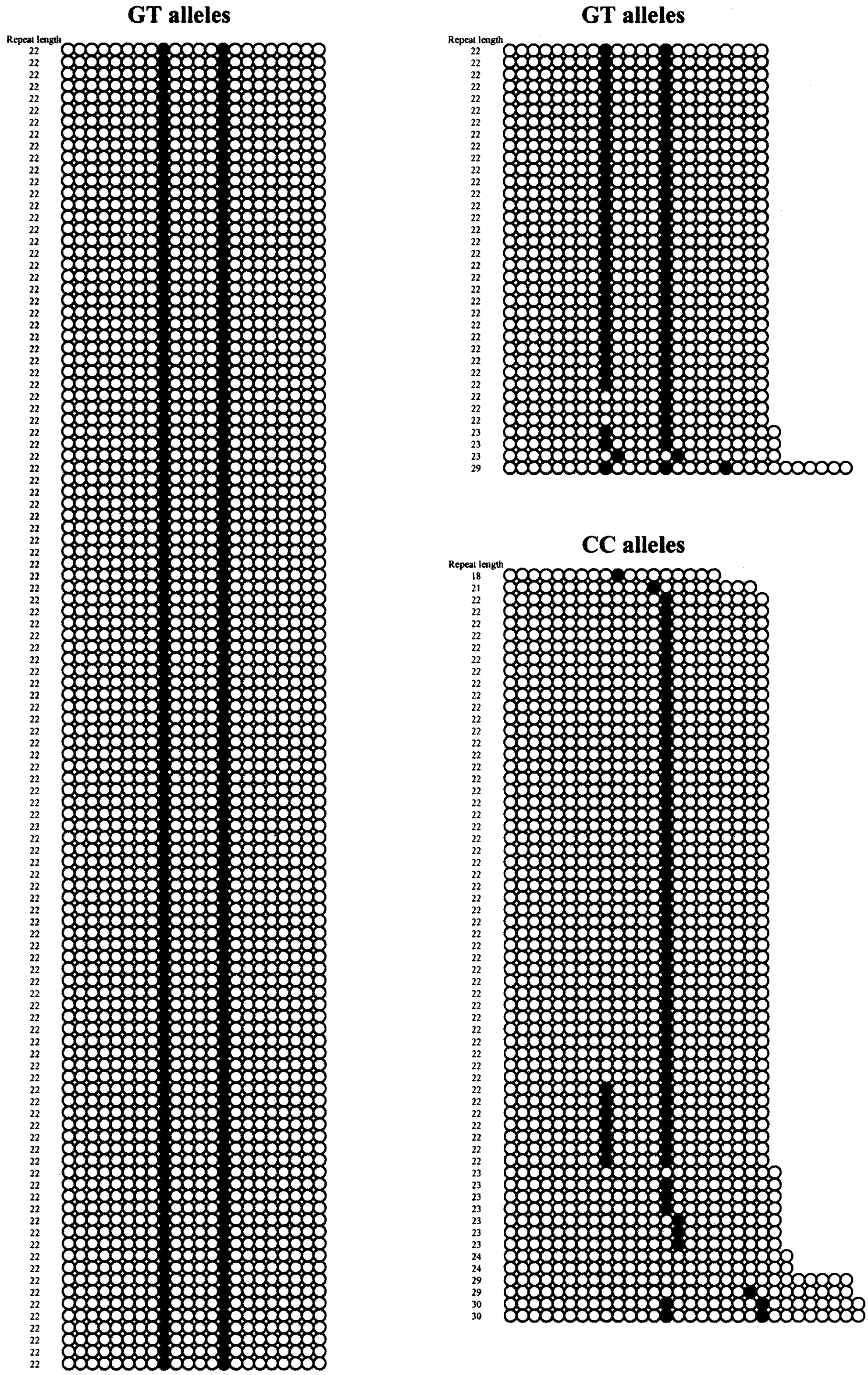


Figure 2. Distribution of CAA triplets in the SCA2 CAG repeat tract of 215 normal chromosomes. CAG repeats are represented by open circles and CAA interruptions are represented by closed circles. Alleles are grouped by SNP haplotype and are arranged in ascending order of the repeat length.

one or were devoid of CAA interruption (Fig. 2). This difference in the number of interruptions present within the CAG repeat tracts of chromosomes with GT and CC haplotype is statistically quite significant ($\chi^2 = 150.122$, $P < 0.0001$). The

Table 3. Linkage disequilibrium between polymorphic markers and SCA2 mutation

Marker	Allele	Patients (frequency)	Controls (frequency)	*P*-value	Association
D12S1333	1	16/17 (0.94)	25/200 (0.13)	0.0000	Positive
	3	1/17 (0.06)	2/200 (0.01)	0.2269	
D12S1672	3	11/17 (0.65)	30/200 (0.15)	0.0013	Positive
	4	5/17 (0.29)	10/200 (0.05)	0.0077	Positive
	5	1/17 (0.06)	12/200 (0.06)	1.0000	
D12S1332	5	7/17 (0.41)	61/200 (0.30)	0.6164	
	6	2/17 (0.12)	25/200 (0.12)	1.0000	
	7	1/17 (0.06)	43/200 (0.22)	0.3248	
	8	7/17 (0.41)	16/200 (0.08)	0.0035	Positive
Truncated haplotypes					
SNP haplotype	CC	17/17 (1.00)	63/215 (0.29)	0.0014	Positive
D12S1333-SNP	1-CC	16/17 (0.94)	6/83 (0.07)	0.0000	Positive
D12S1333-SNP-D12S1672	1-CC-3	11/17 (0.65)	2/83 (0.02)	0.0000	Positive
D12S1333-SNP-D12S1672-D12S1332	1-CC-3-8	6/17 (0.35)	1/83 (0.01)	0.0003	Positive

Association between marker alleles and SCA2 mutation was determined by Fisher's exact test. Alleles (or haplotypes) showing linkage disequilibrium with SCA2 mutation are indicated as positive (*P*-value < 0.01).

Table 4. Associations between microsatellite alleles and SNP haplotypes

Marker	Allele	Controls with GT haplotype (frequency)	Controls with CC haplotype (frequency)	*P*-value
D12S1333	**1**	**10/141 (0.07)**	**15/59 (0.25)**	**0.0033**
	3	1/141 (0.007)	1/59 (0.02)	0.5069
D12S1672	**3**	**4/141 (0.03)**	**26/59 (0.44)**	**0.0000**
	4	**3/141 (0.02)**	**6/59 (0.10)**	**0.0274**
	5	2/141 (0.01)	10/59 (0.17)	0.0003
D12S1332	5	45/141 (0.32)	16/59 (0.27)	0.7468
	6	19/141 (0.13)	6/59 (0.10)	0.6466
	7	28/141 (0.20)	15/59 (0.25)	0.4721
	8	**12/141 (0.08)**	**4/59 (0.07)**	**0.7507**

The frequency distribution of normal control chromosomes with GT and CC haplotypes for the microsatellite marker alleles associated with expanded chromosomes in SCA2 pedigrees (Table 3). The alleles, which showed strong linkage disequilibrium with SCA2 expansion mutation, are indicated in bold and are discussed in the text.

first 5′ CAA interruption was observed at the triplet position 9 and the second at position 14 in 97.4% (148/152) of the GT chromosomes. In contrast, 73.0% (46/63) of the chromosomes with CC haplotype had their first 5′ interruption at position 14 indicating the absence of the most proximal 5′ CAA interruption.

Polar variation within the SCA2 repeat

Based on the most common SCA2 repeat substructure $(CAG)_8CAA(CAG)_4CAA(CAG)_8$, the SCA2 repeat stretch was divided into three continuous CAG repeat tracts: a 5′ CAG repeat tract proximal to the first interruption, a middle tract bordered by two interruptions and a 3′ tract, distal to the last interruption. For alleles with a single CAA interruption, the number of uninterrupted CAG repeat tracts was reduced to two: a 5′ and a 3′ portion. When the number of interruptions was greater than two, alleles were considered to possess two or more middle tracts of CAG repeats.

In order to determine which of these repeat tracts display the length polymorphism seen in normal chromosomes, we sequence analyzed SCA2 normal chromosomes containing a spectrum of repeat sizes. A total of 22 distinct allele types were identified based on repeat length and CAA interruption pattern (Fig. 3). A continuous, uninterrupted CAG repeat configuration was observed on six allele types with CAG repeat length ranging from 16 to 29 and all having CC haplotype. For the remaining 16 interrupted allele types, the 5′ tract of pure repeats was much more variable (range 5–22 repeats) than either the middle (range 4–7) or the 3′ tract (range 8–13 repeats), indicating that for different size SCA2 alleles the vast majority of repeat length variability has occurred at the 5′ end of the repeat (Fig. 3).

Origin of the SNP alleles

To further determine the ancestral status of the two SNPs and to compare the SCA2 CAG repeat organization, we tested them in chimpanzee (GenBank accession no. AF330028; www.ncbi.nlm.nih.gov/Genbank/), gorilla (accession no. AF330029) and various Old World monkeys: langur (accession no. AF330030), rhesus monkey (accession no. AF330031), baboon (accession no. AF330032) and bonnet macaque (accession no. AF330033). The region containing the SNPs and the CAG repeat tract was conserved in evolution. The primer set used to detect these polymorphic sites in the human genome also amplified comparative DNA fragments

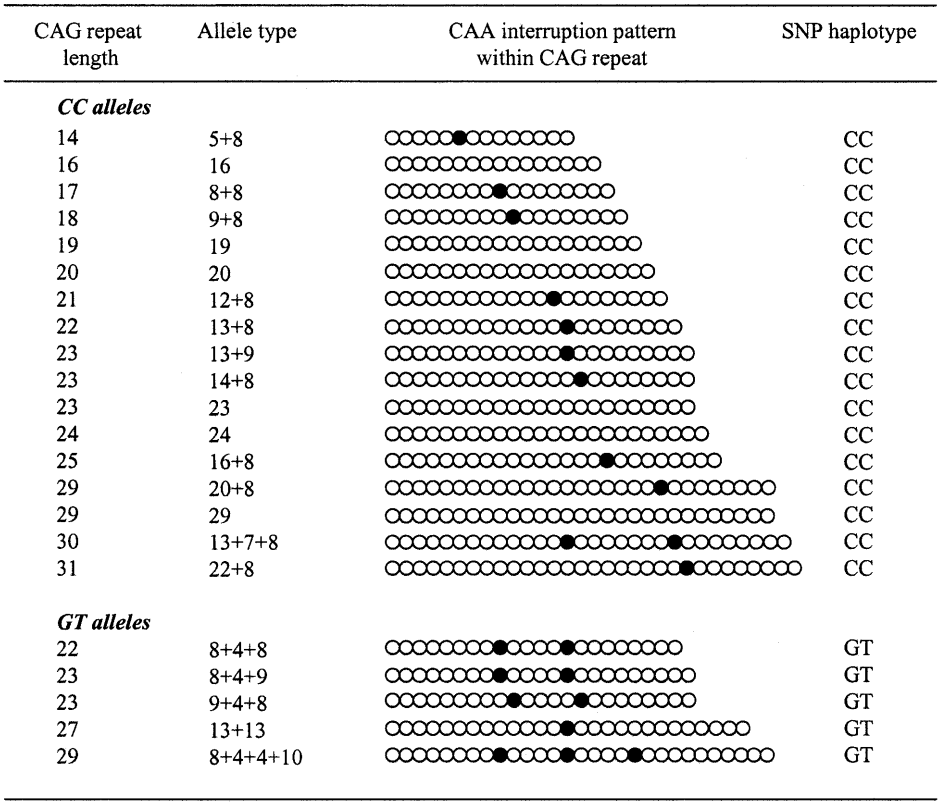


Figure 3. CAA interruption pattern of various lengths of SCA2 repeat. Open circles represent CAG triplets while closed circles represent CAA interruptions. Alleles are grouped by SNP haplotype and are arranged by increasing overall repeat length.

from the genomes of other non-human primates. The CAG repeat length, the repeat substructure and the SNP haplotype of all the samples analyzed is shown in Figure 4. The CAG repeat tracts were not only polymorphic in length but were also interrupted by varying number of CAA triplets. The CCG triplets adjacent to the CAG repeats were also found to be polymorphic and their number varied from one to six in various species examined. All the samples studied had CC haplotype (Fig. 4), suggesting that CC represents the ancestral form in mammalian evolution.

DISCUSSION

We have identified and characterized two novel SNPs in exon 1 of the human SCA2 gene. Two haplotypes distinguished by these two SNP markers, GT and CC, were represented in 70.7 and 29.3% of the normal SCA2 chromosomes, respectively. In contrast, all the disease chromosomes examined in our sample set shared a common CC haplotype. Haplotype studies performed using these SNP markers and linked microsatellites indicated major founder effects in the origin of expanded SCA2 alleles in the Indian population, with a subset of normal chromosomes with CC haplotype predisposed to undergo subsequent expansion. This linkage disequilibrium of SCA2 expanded alleles to CC haplotype suggests that these SNPs are likely to be linked to *cis* acting factors that directly influence repeat stability. Therefore, we used these SNPs to examine various genetic factors, such as total repeat length, CAA interruption pattern and haplotype background, which might contribute to repeat instability and expansion at SCA2 locus.

Sequence analysis of SCA2 CAG repeat for the presence of interrupting CAA triplets revealed that 98% of the normal alleles are interrupted by one or more CAA interruptions, whereas all the expanded alleles contain uninterrupted CAG repeats. Ninety-eight percent of the normal chromosomes with GT haplotype had two or more CAA interruptions with 8+4+8 as the common interspersion pattern. Interestingly, the CC haplotype seen on 100% of the expanded chromosomes was present on normal chromosomes, 86% having single or no CAA interruptions and 13+8 as the most common repeat configuration. This data suggests that the chromosomes with CC haplotype either lack the most proximal CAA interruption or have a tendency for recurrent loss of 5′ CAA triplet by a mechanism which tends to preserve the overall length of the repeat. It has been postulated that a minimal length of pure repeats is required to initiate instability at a repeat locus. The presence of interruptions breaks the repeat into smaller repeat tracts and thus protects it from instability by reducing the length of continuous uninterrupted repeats. There is evidence in the case of SCA1 and fragile X syndrome that larger uninterrupted repeats are more likely to expand than cryptic repeats (21–24). This is also true for dinucleotide repeats where the degree of polymorphism for a repeat locus is generally proportional to the length of the perfect repeat (20). The protective mechanism of interruptions may very well be due to their ability to impair slippage between the complementary strands

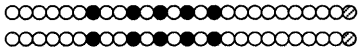
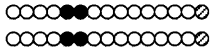
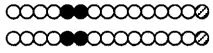
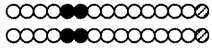
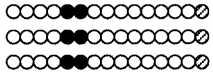
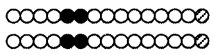
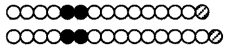
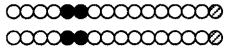
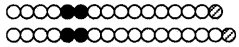
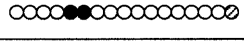
Primates	CAG repeat length	Repeat substructure	SNP haplotype
Chimpanzee:			
1.....	27		CC
2.....	27		CC
Gorilla:			
1.....	18		CC
2.....	18		CC
Langur:			
1.....	24		CC
2.....	25		CC
3.....	25		CC
4.....	25		CC
Baboon:			
1.....	15		CC
2.....	15		CC
3.....	15		CC
4.....	19		CC
Rhesus Monkey:			
1.....	16		CC
2.....	16		CC
3.....	16		CC
4.....	17		CC
Bonnet Macaque:			
1.....	14		CC
2.....	14		CC
3.....	14		CC
4.....	14		CC
5.....	14		CC
6.....	14		CC
7.....	14		CC
8.....	14		CC
9.....	14		CC
10.....	14		CC
11.....	14		CC
12.....	14		CC
13.....	14		CC
14.....	14		CC
15.....	14		CC
16.....	14		CC
17.....	15		CC
18.....	15		CC
19.....	16		CC
20.....	16		CC

Figure 4. CAG repeat length, repeat substructure and SNP haplotype in non-human primates. Open circles represent CAG triplets; dark circles represent CAA interruptions and hatched circles represent CCG triplets adjacent to the CAG repeat tract.

(25). Thus, absence of 5′ CAA interruption in chromosomes with CC haplotype might represent one of the factors contributing to repeat instability presumably because only a single event will be required to create a perfect CAG repeat and hence predispose to expansion. A recent report by Costanzi-Porrini *et al.* (29) of two patients with SCA2 phenotype and having an interrupted 34 CAG repeat allele with configuration 24+8, suggests that the loss of all interruptions is not absolutely necessary for repeat expansion. A slow sequential lengthening of 5′ repeat tract in alleles lacking the most proximal CAA interruption can also result in expansion to pathological range at the SCA2 locus. Therefore, predisposition towards repeat instability and expansion in SCA2 appears correlated not strictly to the repeat length but rather to the purity of the repeat.

The polar variation in SCA2 repeat is similar to that observed for CGG repeats in fragile X syndrome and in hypervariable human and murine minisatellite loci where allele diversity is largely determined by sequence changes at one end of the repeat array (22,23,30). Unlike fragile X syndrome where the majority of the changes occur in the most 3′ tract (22,23), the 5′ tract (relative to the orientation of transcription) of SCA2 CAG repeat was found to be more variable. The observed polar variation among SCA2 alleles demonstrates that the replication slippage is a frequent event in the 5′ portion

Table 5. SCA2 repeat interruption patterns

Repeat length	Repeat substructure	No. of CAAs	No. of chromosomes (frequency)
18	9+8	1	1 (0.005)
21	12+8	1	1 (0.005)
22	8+4+8	2	152 (0.707)
22	13+8	1	44 (0.205)
23	9+4+8	2	1 (0.005)
23	8+4+9	2	2 (0.009)
23	13+9	1	3 (0.014)
23	14+8	1	3 (0.014)
23	23	0	1 (0.005)
24	24	0	2 (0.009)
29	8+4+4+10	3	1 (0.005)
29	20+8	1	1 (0.005)
29	29	0	1 (0.005)
30	13+7+8	2	2 (0.009)
Total number of chromosomes (*n*)			215

The interruption patterns of SCA2 CAG repeat alleles and the frequency of their occurrence among 215 randomly selected normal chromosomes. CAA interruption patterns have been abbreviated. A plus sign represents a CAA and the number denotes the uninterrupted CAG repeats (Materials and Methods). The alleles are arranged in ascending order of repeat length.

of the CAG repeat and supports the hypothesis that it is the lack of the most proximal 5′ CAA interruption that acts to destabilize the CAG repeat region and results in the highest variability of pure repeat lengths at 5′ end. The observed variability within the continuous tract of CAG repeats and the fact that changes involve differences in multiples of 3 bp clearly favors slipped strand mispairing as the most likely mechanism of repeat length variability at the SCA2 locus (31). Unequal sister chromatid exchange and gene conversions have also been proposed as a mechanism of repeat length variability (32). However, if these were the dominant mode of repeat length variation for SCA2 repeats then one might expect larger alleles to contain three or more CAAs as they are constructed in a cassette like arrangement by conversion or unequal cross over. Such alleles have been observed but appear to be rare (Table 5 and Fig. 2).

Analysis of comparative DNA fragments from the genome of other non-human primates gave insight into the evolutionary history of SCA2 CAG repeat with CC haplotype representing the ancestral form in mammalian evolution. The absence of two other haplotypes (GC and CT) among the four possible combinations (GT, CC, GC and CT) suggests that the mutation from C to G and C to T creating the GT allele might have occurred simultaneously and this new haplotype GT, either through selection (repeat stability) or genetic drift, became the more common allele associated with normal SCA2 chromosomes in our population.

In summary, we report here a very comprehensive study of the factors that may drive variability and ultimately repeat instability at the SCA2 locus. Two newly characterized SNPs, the closest of the surrounding markers to the SCA2 CAG repeat, showed a complete association of expansion mutation with CC haplotype in our population. The identification of a common CC haplotype for disease chromosomes facilitated the characterization of the factors involved in repeat instability at the SCA2 locus. Based on our data we demonstrate that CAA interruptions play a pivotal role in restricting mutability at the SCA2 locus and that their absence predisposes the alleles to expansion. However, larger surveys of SCA2 repeat substructure and SNP haplotype in different ethnic groups will be necessary to further confirm these findings.

MATERIALS AND METHODS

Subjects

The study was carried out on 22 SCA2 pedigrees comprising of 123 members including patients and family members. Some of these pedigrees have been reported previously (16). To establish the distribution of SNPs and CAA interruption pattern, 215 normal chromosomes were analyzed as controls. These chromosomes were selected to be representative of overall distribution of repeat length that we have observed for the SCA2 locus (16). All the subjects were of Indian origin and mixed geographical derivation. Since only ~8% of the normal SCA2 alleles possess other than 22 repeats, an additional 15 alleles with repeat number other than 22 were also analyzed. Blood samples were collected from all patients and normal individuals with informed consent. Ten apparently unrelated bonnet macaques (*Macaca radiata*), two baboons (*Papio hamadryas*), two rhesus monkeys (*Macaca mulatta*), two langurs (*Presbytis entellus*), a gorilla (*Gorilla gorilla*) and a chimpanzee (*Pan troglodytes*) sample were also investigated.

Amplification of SCA2 CAG repeat region

Genomic DNA was isolated from peripheral blood leukocytes using a modification of the salting out procedure (33). The region containing the SCA2 CAG repeat was PCR amplified using previously published primers (3) and the exact size of the repeat in the fluorescently labeled PCR product was determined by Gene Scan analysis using an ABI Prism 377 automated DNA sequencer (Perkin Elmer, Foster City, CA).

SNP detection and CAA interspersion analysis

Sequencing analysis was carried out for characterization of two SNPs and CAA interruption pattern of SCA2 repeat. The region containing the SNPs and the CAG repeat stretch was amplified using primers SCA2-FP3 (5′-CTCCGCCTCAGAC-TGTTTTGGTAG-3′) and SCA2-RP3 (5′-GTGGCCGAG GACGAGGAGAC-3′). Approximately 100 ng of genomic DNA was amplified in a 50 µl reaction volume containing a final concentration of 5 mM Tris, 25 mM KCl, 0.75 mM $MgCl_2$, 0.05% gelatin, 20 pmol of each primer, 200 µM dNTPs and 0.5 U of *Taq* DNA polymerase. Samples were denatured at 94°C for 3 min followed by 35 cycles of denaturation (94°C, 45 s), annealing (52°C, 30 s), extension (72°C, 45 s) and a final extension of 7 min at 72°C in a Perkin Elmer GeneAmp PCR System 9600. The PCR products were purified from bands cut out of agarose gel using a QIAquick gel extraction kit (Qiagen Inc., CA) and were directly sequenced using dye

terminator chemistry on an ABI Prism 377 automated DNA sequencer with the PCR primers. For SCA2 positive and normal control families, the phase of the two SNP alleles was determined based on the segregation pattern. For normal control individuals when sample showed homozygosity, phase was easily assigned. For heterozygous samples, the phases could be determined by performing sequencing on the parents' DNA samples and deducing the phase based on the segregation pattern. The phases could not be resolved unambiguously for some samples and were excluded from the study.

Nomenclature of CAG repeat configuration

CAG repeat interruption patterns are summarized as follows: a plus sign (+) designates the position of a CAA interruption and the number refers to the length of uninterrupted CAG repeats. A 8+4+8 allele, for example, symbolizes the sequence of $(CAG)_8CAA(CAG)_4CAA(CAG)_8$. CAA interruptions are described in the text as 5′ or 3′ relative to the orientation of transcription of the SCA2 gene.

Haplotype analysis

Haplotypes were generated for the SCA2 positive families using three microsatellite markers D12S1333, D12S1672 and D12S1332. These markers span a region around the SCA2 CAG repeat in the following order: telomere-D12S1333-200 kb-D12S1672-350 kb-D12S1332-centromere, with D12S1333 and D12S1332 flanking the CAG repeat (Fig. 1). D12S1672 is in the first intron of the SCA2 gene and is 20 kb centromeric to the CAG repeat stretch, which is in the first exon of the SCA2 gene (3,34). The frequency of these markers was also determined in the normal control chromosomes. In all the analyses the CEPH family member 1347-02 was used as a control for verification of repeat size.

ACKNOWLEDGEMENTS

We thank Dr Nitai P.Bhattacharyya and Dr Partha P.Majumder for their generous contribution of few normal and SCA2 samples for this study. We are grateful to Dr Q.Saleem for helpful discussions and Dr C.B.Rao for assistance with statistical analysis, Ms R.Jaya, Ms M.Ruchi and Ms T.Sakshi for help with Genescan and sequence analysis, and Ms Manju, Ms Gurjit and Ms Anuradha for technical support. We would like to acknowledge Ms Sanghamitra Roy for extraction of patient DNA samples. We would also like to thank the Primate Research Facilities of the National Institute of Immunology, New Delhi, the Indian Institute of Science, Bangalore and Dr Lalji Singh of the Centre for Cellular and Molecular Biology, Hyderabad, for providing the primate samples. Financial support from the Department of Biotechnology, Government of India to the Program on Functional Genomics to S.K.B. is duly acknowledged.

REFERENCES

1. Orr, H.T., Chung, M.Y., Banfi, S., Kwiatkowski, T.J.,Jr, Servadio, A., Beaudet, A.L., McCall, A.E., Duvick, L.A., Ranum, L.P. and Zoghbi, H.Y. (1993) Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. *Nat. Genet.*, **4**, 221–226.
2. Imbert, G., Saudou, F., Yvert, G., Devys, D., Trottier, Y., Garnier, J.M., Weber, C., Mandel, J.L., Cancel, G., Abbas, N. *et al.* (1996) Cloning of the gene for spinocerebellar ataxia 2 reveals a locus with high sensitivity to expanded CAG/glutamine repeats. *Nat. Genet.*, **14**, 285–291.
3. Pulst, S.M., Nechiporuk, A., Nechiporuk, T., Gispert, S., Chen, X.N., Lopes-Cendes, I., Pearlman, S., Starkman, S., Orozco-Diaz, G., Lunkes, A. *et al.* (1996) Moderate expansion of a normally biallelic trinucleotide repeat in spinocerebellar ataxia type 2. *Nat. Genet.*, **4**, 269–276.
4. Sanpei, K., Takano, H., Igarashi, S., Sato, T., Oyake, M., Sasaki, H., Wakisaka, A., Tashiro, K., Ishida, Y., Ikeuchi, T. *et al.* (1996) Identification of the spinocerebellar ataxia type 2 gene using a direct identification of repeat expansion and cloning technique, DIRECT. *Nat. Genet.*, **14**, 277–284.
5. Kawaguchi, Y., Okamoto, T., Taniwaki, M., Aizawa, M., Inoue, M., Katayama, S., Kawakami, H., Nakamura, S., Nishimura, M., Akiguchi, I. *et al.* (1994) CAG expansion in a novel gene for Machado-Joseph disease at chromosome 14q32.1. *Nat. Genet.*, **8**, 221–227.
6. Zhuchenko, O., Bailey, J., Bonnen, P., Ashizawa, T., Stockton, D.W., Amos, C., Dobyns, W.B., Subramony, S.H., Zoghbi, H.Y. and Lee, C.C. (1997) Autosomal dominant cerebellar ataxia (SCA6) associated with small polyglutamine expansions in the alpha 1A-voltage-dependent calcium channel. *Nat. Genet.*, **15**, 62–69.
7. David, G., Abbas, N., Stevanin, G., Durr, A., Yvert, G., Cancel, G., Weber, C., Imbert, G., Saudou, F., Antoniou, E. *et al.* (1997) Cloning of the SCA7 gene reveals a highly unstable CAG repeat expansion. *Nat. Genet.*, **17**, 65–70.
8. Koob, M.D., Moseley, M.L., Schut, L.J., Benzow, K.A., Bird, T.D., Day, J.W. and Ranum, L.P. (1999) An untranslated CTG expansion causes a novel form of spinocerebellar ataxia (SCA8). *Nat. Genet.*, **21**, 379–384.
9. Matsuura, T., Yamagata, T., Burgess, D.L., Rasmussen, A., Grewal, R.P., Watase, K., Khajavi, M., McCall, A.E., Davis, C.F., Zu, L. *et al.* (2000) Large expansion of the ATTCT pentanucleotide repeat in spinocerebellar ataxia type 10. *Nat. Genet.*, **26**, 191–194.
10. Holmes, S.E., O'Hearn, E.E., McInnis, M.G., Gorelick-Feldman, D.A., Kleiderlein, J.J., Callahan, C., Kwak, N.G., Ingersoll-Ashworth, R.G., Sherr, M., Sumner, A.J. *et al.* (1999) Expansion of a novel CAG trinucleotide repeat in the 5′ region of PPP2R2B is associated with SCA12. *Nat. Genet.*, **23**, 391–392.
11. Gispert, S., Twells, R., Orozco, G., Brice, A., Weber, J., Heredero, L., Scheufler, K., Riley, B., Allotey, R., Nothers, C. *et al.* (1993) Chromosomal assignment of the second locus for autosomal dominant cerebellar ataxia (SCA2) to chromosome 12q23–24.1. *Nat. Genet.*, **4**, 295–299.
12. Cancel, G., Durr, A., Didierjean, O., Imbert, G., Burk, K., Lezin, A., Belal, S., Benomar, A., Abada-Bendib, M., Vial, C. *et al.* (1997) Molecular and clinical correlations in spinocerebellar ataxia 2: a study of 32 families. *Hum. Mol. Genet.*, **6**, 709–715.
13. Geschwind, D.H., Perlman, S., Figueroa, C.P., Treiman, L.J. and Pulst, S.M. (1997) The prevalence and wide clinical spectrum of the spinocerebellar ataxia type 2 trinucleotide repeat in patients with autosomal dominant cerebellar ataxia. *Am. J. Hum. Genet.*, **60**, 842–850.
14. Lorenzetti, D., Bohlega, S. and Zoghbi, H.Y. (1997) The expansion of the CAG repeat in ataxin-2 is a frequent cause of autosomal dominant spinocerebellar ataxia. *Neurology*, **49**, 1009–1013.
15. Giunti, P., Sabbadini, G., Sweeney, M.G., Davis, M.B., Veneziano, L., Mantuano, E., Federico, A., Plasmati, R., Frontali, M. and Wood, N.W. (1998) The role of the SCA2 trinucleotide repeat expansion in 89 autosomal dominant cerebellar ataxia families. Frequency, clinical and genetic correlates. *Brain*, **121**, 459–467.
16. Saleem, Q., Choudhry, S., Mukerji, M., Bashyam, L., Padma, M.V., Chakravarthy, A., Maheshwari, M.C., Jain, S. and Brahmachari, S.K. (2000) Molecular analysis of autosomal dominant hereditary ataxias in the Indian population: high frequency of SCA2 and evidence for a common founder mutation. *Hum. Genet.*, **106**, 179–187.
17. Weber, J.L. (1990) Informativeness of human (dC-dA)n.(dG-dT)n polymorphisms. *Genomics*, **7**, 524–530.
18. Bhandari, R. and Brahmachari, S.K. (1995) Analysis of CAG/CTG triplet repeats in the human genome: Implication of transcription factor gene regulation. *J. Biosci.*, **5**, 613–627.
19. Zoghbi, H.Y. (1996) The expanding world of ataxins. *Nat. Genet.*, **14**, 237–238.
20. Wells, R.D. and Warren, S.T. (1998) *Genetic Instability and Hereditary Neurological Disease*. Academic Press, San Diego, CA.
21. Chung, M.Y., Ranum, L.P., Duvick, L.A., Servadio, A., Zoghbi, H.Y. and Orr, H.T. (1993) Evidence for a mechanism predisposing to intergenerational CAG repeat instability in spinocerebellar ataxia type I. *Nat. Genet.*, **5**, 254–258.

22. Kunst, C.B. and Warren, S.T. (1994) Cryptic and polar variation of the fragile X repeat could result in predisposing normal alleles. *Cell*, **77**, 853–861.
23. Eichler, E.E., Holden, J.J., Popovich, B.W., Reiss, A.L., Snow, K., Thibodeau, S.N., Richards, C.S., Ward, P.A. and Nelson, D.L. (1994) Length of uninterrupted CGG repeats determines instability in the FMR1 gene. *Nat. Genet.*, **8**, 88–94.
24. Zhong, N., Yang, W., Dobkin, C. and Brown, W.T. (1995) Fragile X gene instability: anchoring AGGs and linked microsatellites. *Am. J. Hum. Genet.*, **57**, 351–361.
25. Pearson, C.E., Eichler, E.E., Lorenzetti, D., Kramer, S.F., Zoghbi, H.Y., Nelson, D.L. and Sinden, R.R. (1998) Interruptions in the triplet repeats of SCA1 and FRAXA reduce the propensity and complexity of slipped strand DNA (S-DNA) formation. *Biochemistry*, **37**, 2701–2708.
26. Basu, P., Chattopadhyay, B., Gangopadhaya, P.K., Mukherjee, S.C., Sinha, K.K., Das, S.K., Roychoudhury, S., Majumder, P.P. and Bhattacharyya, N.P. (2000) Analysis of CAG repeats in SCA1, SCA2, SCA3, SCA6, SCA7 and DRPLA loci in spinocerebellar ataxia patients and distribution of CAG repeats at the SCA1, SCA2 and SCA6 loci in nine ethnic populations of eastern India. *Hum. Genet.*, **106**, 597–604.
27. Wadia, N., Pang, J., Desai, J., Mankodi, A., Desai, M. and Chamberlain, S. (1998) A clinicogenetic analysis of six Indian spinocerebellar ataxia (SCA2) pedigrees. The significance of slow saccades in diagnosis. *Brain*, **121**, 2341–2355.
28. Pang, J., Allotey, R., Wadia, N., Sasaki, H., Bindoff, L. and Chamberlain, S. (1999) A common disease haplotype segregating in spinocerebellar ataxia 2 (SCA2) pedigrees of diverse ethnic origin. *Eur. J. Hum. Genet.*, **7**, 841–845.
29. Costanzi-Porrini, S., Tessarolo, D., Abbruzzese, C., Liguori, M., Ashizawa, T. and Giacanelli, M. (2000) An interrupted 34-CAG repeat SCA-2 allele in patients with sporadic spinocerebellar ataxia. *Neurology*, **54**, 491–493.
30. Armour, J.A., Harris, P.C. and Jeffreys, A.J. (1993) Allelic diversity at minisatellite MS205 (D16S309): evidence for polarized variability. *Hum. Mol. Genet.*, **2**, 1137–1145.
31. Pearson, C.E. and Sinden, R.R. (1996) Alternative structures in duplex DNA formed within the trinucleotide repeats of the myotonic dystrophy and fragile X loci. *Biochemistry*, **35**, 5041–5053.
32. Jeffreys, A.J., Tamaki, K., MacLeod, A., Monckton, D.G., Neil, D.L. and Armour, J.A. (1994) Complex gene conversion events in germline mutation at human minisatellites. *Nat. Genet.*, **6**, 136–145.
33. Miller, S.A., Dykes, D.D. and Polesky, H.F. (1988) A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res.*, **16**, 1215.
34. Sahba, S., Nechiporuk, A., Figueroa, K.P., Nechiporuk, T. and Pulst, S.M. (1998) Genomic structure of the human gene for spinocerebellar ataxia type 2 (SCA2) on chromosome 12q24.1. *Genomics*, **47**, 359–364.