Human Molecular Genetics, 1999, Vol. 8, No. 6 **1061–1067**

Cis-acting modifiers of expanded CAG/CTG triplet repeat expandability: associations with flanking GC content and proximity to CpG islands

Graham J. R. Brock, Niall H. Anderson[1] and Darren G. Monckton[+]

Division of Molecular Genetics, Institute of Biomedical and Life Sciences, University of Glasgow, Anderson College, 56 Dumbarton Road, Glasgow G11 6NU, UK and [1]Department of Medicine and Therapeutics, Western Infirmary, University of Glasgow, Glasgow G11 6NT, UK

Received January 11, 1999; Revised and Accepted March 2, 1999

An increasing number of human genetic disorders are associated with the expansion of trinucleotide repeats. The majority of these diseases are associated with CAG/CTG expansions, including Huntington's disease, myotonic dystrophy and many of the spinocerebellar ataxias. Recently, two new expanded CAG/CTG repeats have been identified that are not associated with a phenotype. Expanded alleles at all of these loci are unstable, with frequent length changes during intergenerational transmission. However, variation in the relative levels of instability, and the size and direction of the length change mutations observed, between the CAG/CTG loci is apparent. We have quantified these differences, taking into account effects of progenitor allele length, by calculating the relative expandability of each repeat. Since the repeat motifs are the same, these differences must be a result of flanking sequence modifiers. We present data that indicate a strong correlation between the relative expandability of these repeats and the flanking GC content. Moreover, we demonstrate that the most expandable loci are all located within CpG islands. These data provide the first insights into the molecular bases of *cis*-acting flanking sequences modifying the relative mutability of dispersed expanded human triplet repeats.

INTRODUCTION

Expanded trinucleotide repeat sequences have been associated with a growing number of inherited human disorders with unusual genetics. The majority characterized so far involve the expansion of the trinucleotide sequence CAG/CTG, including the spinocerebellar ataxias, type 1 (SCA1) (1), type 2 (SCA2) (2), type 3 (SCA3, also known as Machado–Joseph disease) (3), type 7 (SCA7) (4), dentatorubral-pallidoluysian atrophy (DRPLA) (5), spino-bulbar muscular atrophy (SBMA) (6), Huntington's disease (HD) (7) and myotonic dystrophy (DM) (8–10). With the exception of the *DM* repeat, all these other CAG/CTG disease-associated repeat expansions occur in translated regions of the gene and encode polyglutamine. The *SCA2*, *SCA3*, *HD* and *SBMA* repeats are found in the first exon, whilst those at the *SCA1*, *SCA7* and *DRPLA* loci are located in the eighth, third and fifth exons, respectively. The *DM* repeat occurs in the 3'-untranslated region of the DM protein kinase gene (*DMPK*) and in the putative promoter of the DM-associated homeodomain gene (*DMAHP*) (11). At all of these loci, the repeats are polymorphic in length in the normal population, with alleles up to ~35 CAGs. Disease-associated alleles are expanded relative to those found in the normal population, with array lengths from >35 up to hundreds, or even thousands of repeats. Two further CAG/CTG expansions (>35 repeats), not yet associated with a phenotype, have also been characterized. One of these, *CTG18.1,* is located in an intron of the SL3-3 enhancer factor 2 gene (*SEF2*) (12). It currently remains unclear whether the other, expanded repeat domain CAG/CTG (*ERDA1*), is associated with a gene, given the conflicting data regarding transcription from the region (13,14).

All of these expanded CAG/CTG repeats have been described as 'dynamic mutations', with repeat number changes frequent during intergenerational transmission. Such intergenerational length changes at the disease loci usually consist of an increase in repeat number, providing a molecular basis for anticipation, in which there is an increase in disease severity and a decrease in age of onset in succeeding generations (15). Repeat number changes are influenced by the sex of the transmitting parent, the number of repeats and the presence of interruptions within the repeat array (16). Usually germline mutations are more frequent and larger when transmitted by a male, and larger alleles have a higher mutation rate and undergo larger length change mutations. In contrast, interruptions within the array are known to stabilize repeat tracts. Normal alleles at the *SCA1* and *SCA2* loci contain interruptions within the array, though expanded alleles consist of pure CAG tracts (17,18). The *SCA3* repeat is preceded by an 18 bp cryptic repeat (CAG CAG CAA AAG CAG CAA) followed by a pure CAG tract. The cryptic repeat is conserved in expanded alleles, but the expanded region comprises pure CAG repeats (3). Similarly, the HD repeat is followed by two cryptic repeats (CAA CAG); although occasional expanded alleles have lost the CAA interruption, increasing the effective array length

---

[+]To whom correspondence should be addressed. Tel: +44 141 330 6213; Fax: +44 141 330 6871; Email: dmonck@molgen.gla.ac.uk

Table 1. CAG/CTG triplet repeat expandability, flanking sequence analyses and genomic localization

Locus	Accession no.	%GC (100 bp flanking)	%GC (500 bp flanking)	Estimated expandability (95% confidence interval)			Chromosomal location	Refs[b]
				Male	Female	Sperm		
DM	X84813 and l00727	69.5	66	4.81 (3.46–7.22)	7.64 (5.16–10.87)	4.34 (4.17–4.50)	19q13.3	(21)
SCA7	AF032102	83.5	71.5	1.30 (0.80–1.65)	0.40 (0.27–0.56)	7.80 (7.52–7.98)	3p12–13	(41–43)
SCA2	AC004085	77	79	0.97 (0.65–1.33)	0.45 (0.25–0.64)	n/a	12q24.1	(18,44)
HD	Z68756	74.5	71	0.29 (0.21–0.43)	0.09 (0.00–0.17)	0.98 (0.84–1.0)	4p16.3	(22,45)
DRPLA	U47924	63.5	66	0.19 (0.14–0.24)	0.04 (–0.05–0.14)	n/a	12p13.31	(46,47)
SCA1	AC002326	66	67.2	0.14 (0.00–0.24)	0.00 (–0.05–0.04)	0.26 (0.19–0.35)	6p22–23	(23)
SBMA	X78592 and M27423	65	59	0.08 (0.00–0.22)	0.00 (0.00–0.00)	0.13 (0.08–0.13)	Xq11.2	(20,48)
SCA3	AJ000501	36.5	38.5[a]	0.07 (0.05–0.09)	0.02 (0.01–0.03)	n/a	14q32.1	(24)
ERDA1	AC004108	38.5	43	–0.01 (–0.07–0.02)	0.00 (–0.03–0.05)	n/a	17q21.3	
CTG18.1	U75701	45	n/a	n/a	n/a	n/a	18q21.1	

[a] Only 435 bp of sequence available on the 3' flank; n/a, not available.
[b] Data from the references shown are those not found in the OMIM database, used to collate the expandability figures.

by two repeats, most expanded alleles comprise a pure CAG expansion 5' of the cryptic repeat (19). The vast majority of normal and expanded alleles at the remaining loci are thought to comprise pure CAG tracts. Even taking into account effects of allele length and repeat interruptions, the mutation rate and the size and direction of the length change mutations observed also appear to vary between loci, with some apparently more mutable and liable to expand to greater lengths than others. We have studied published data and quantified the relative expandability at each of the expanded CAG/CTG loci. As the loci share a common repeat motif, possible correlations between the relative expandability and the flanking DNA regions were investigated.

RESULTS

Locus-specific variation of repeat stability

In order to quantify expanded (>35) repeat stability, male and female intergenerational transmissions, as determined by pedigree analysis and previously reported in the literature, were collated for nine of the CAG/CTG loci. It is known that there is a strong correlation between progenitor allele length, and mutation rate and intergenerational length change, with larger alleles usually giving rise to more frequent and greater expansions from one generation to the next (16). However, we noted that for each locus, the progenitor allele length distributions were non-uniform. For instance, progenitor allele lengths ranged from 36 to 44 repeats at the *SCA2* locus and from 62 to 73 repeats at the *SCA3* locus. Thus, in order to quantify the relative level of stability at each locus, we needed to calculate a measure of mutability that was independent of progenitor allele length effects. Given the difference in allele length distribution at each locus, a simple measure of the mutation rate would therefore not suffice. Since the loci appear to differ not only in the rate of the length change mutation events but also in their magnitude, we sought also to incorporate this dimension of the mutation process into a quantitative relative measure of the mutability. In order to account for allele length effects, it would seem reasonable to calculate the relative length change mutation of a given transmission event as a proportion of the progenitor allele length. However, given that 35 repeats is the lower boundary between stable normal alleles and unstable expanded alleles, we have assumed a basal average allele length change of 0 at 35 repeats. Thus, in order to quantify the relative level of mutability at each locus independently of progenitor allele length effects, we have calculated the 'relative length change per expanded progenitor repeat' [i.e. length change/(progenitor allele length – 35 repeats)]. This figure can be described more simply as the 'expandability', and reflects the tendency of an above threshold repeat block to undergo further expansion.

The median relative length change per expanded progenitor repeat at each locus was calculated for both male and female transmissions (Table 1, Fig. 1). These data clearly demonstrate that significant differences in the relative germline expandability exist between many of the loci. In both sexes, *DM* is revealed to be the most expandable locus, followed by *SCA7*, *SCA2* and *HD*. At the remaining loci, the median expandability varies between the sexes, with *SBMA*, *ERDA1* and *SCA1* all showing a median length change of zero in female transmissions. Mutations at the same three loci in male transmissions show a median increase of length change in *SBMA* and *SCA1*, and a decrease of repeats in *ERDA1*. With the exception of *DM* and *ERDA1*, the relative expandability is greater in male transmissions than in females. The sex-averaged values retain the same order as observed in the male transmissions (Fig. 1). At present, insufficient data are available to quantify the relative mutability of the *CTG18.1* locus. However, preliminary sperm analyses indicate that although small length change mutants are common, large length change mutations are rare, suggesting that *CTG18.1* is one of the relatively less expandable loci (F.K. Gould and D.G. Monckton, unpublished data).

Repeat stability in sperm

For some of these loci, repeat length variation in the male germline has also been assessed directly by sperm analysis (20–24) (D.G. Monckton, manuscript in preparation). These

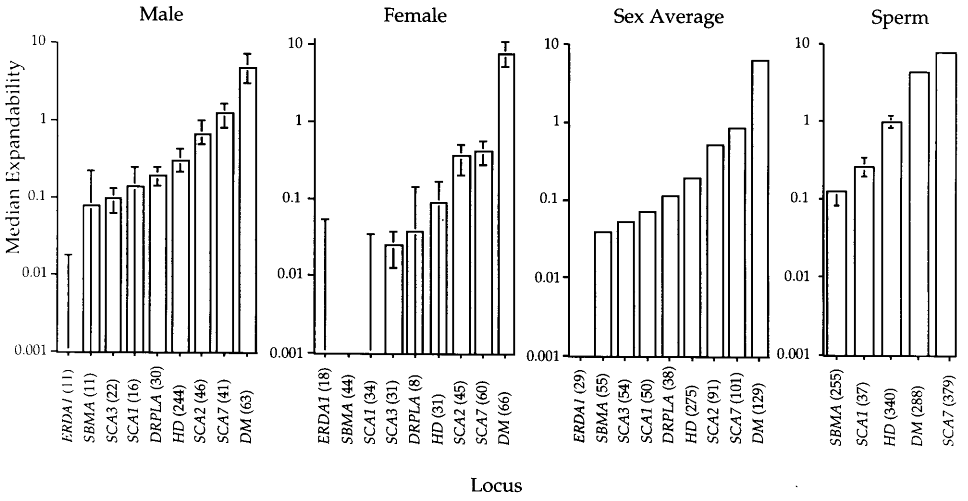

Figure 1. Median relative length change per expanded progenitor repeat. The median expandability in repeat transmissions (relative length change per expanded progenitor repeat) at each locus was calculated as detailed in the text. The expandability (±95% confidence limits) at each locus was plotted on a log scale with the number of transmissions analysed shown in parentheses.

data were analysed in the same way, with the progenitor size being corrected and the median length changes calculated (Table 1, Fig. 1). Relative levels of the CAG/CTG repeat expandability in sperm show a pattern similar to that seen in male and female transmissions. The most expandable loci were again *SCA7*, *DM* and *HD*, with *SCA1* and *SBMA* showing much lower levels of expandability. Interestingly, the sperm data indicate that *SCA7* appears to be more expandable in the male germline than the *DM* repeat (D.G. Monckton, manuscript in preparation). The reported data available for mutability in sperm at the *SCA3* locus did not include analysis of individual sperm (24). It was therefore not possible to calculate a figure for the median relative length change per expanded progenitor repeat. However, the data indicated that the repeat associated with *SCA3* resulted in a mean decrease in transmitted allele in the male germline, confirming it as the least expandable of the loci assayed by direct sperm analysis.

Analysis of flanking DNA sequences

Having determined that significant differences in expandability, independent of repeat length, do indeed exist between the loci, we concluded that this variation must be associated with the surrounding DNA sequences. The genomic locations of the various repeat loci (Table 1) did not reveal any obvious patterns. Thus, we investigated the immediate flanking DNA sequences to determine whether they may influence repeat mutability. Of the loci analysed, *SCA1*, *ERDA1*, *DRPLA*, *HD* and *DM* were at the time of the analyses located within reasonably well-defined genomic sequence contexts with >5 kb of sequence available on both flanks. *SCA2* and *SBMA* were less well characterized, with <5 kb available on the 3' flank. *CTG18.1*, *SCA7* and *SCA3* were even less well characterized, with <2.2 kb of flanking sequence available in total. Analysis of these flanking regions revealed considerable differences in %GC and $\mathrm{CpG}_{\mathrm{obs/exp}}$ (25) between the loci (Fig. 2). For instance, the *ERDA1* repeat is located in a region with relatively low %GC and $\mathrm{CpG}_{\mathrm{obs/exp}}$ (42% and 0.13, respectively, over 10 kb), similar to that of bulk genomic DNA (25). Moreover, analysis of 92 kb of *ERDA1* flanking DNA does not reveal the presence of any known genes or homologies to any other coding sequences in the available databases (data not shown). Although only a limited amount of sequence was available, it suggested that *CTG18.1* and *SCA3* were located in similarly GC-poor and low $\mathrm{CpG}_{\mathrm{obs/exp}}$ regions. In contrast, the *DM* repeat lies in an extremely GC- (63% over 10 kb) and gene-rich region of chromosome 19. Moreover, the $\mathrm{CpG}_{\mathrm{obs/exp}}$ frequency is also very high (0.56 over 10 kb), indicating that as reported previously (11), the *DM* repeat lies in a very large CpG island (CGI). Similarly, the *HD*, *SCA2* and *SCA7* repeats also lie within regions predicted to be CGIs. Small islands were also predicted in the sequences flanking *DRPLA*, *SCA1* and *SBMA*, but these did not include the CAG repeat. The repeats at the *ERDA1*, *CTG18.1* and *SCA3* loci do not lie in a CGI, and none was predicted in the available flanking sequence. In a more detailed analysis, we aligned the 100 bp surrounding each repeat (data not shown). Although no conserved sequence motifs were detected, the levels of %GC and $\mathrm{CpG}_{\mathrm{obs/exp}}$ varied greatly (Table 1).

Correlations between relative mutability and flanking sequences

Having determined that the relative expandability and flanking GC content of the loci differed, we sought to determine whether these variables might be associated. Notably, the four most expandable loci were located within CGIs, whilst the remainder were not, a highly significant association ($P < 0.01$, Fisher's exact test). Moreover, ranked correlations showed that the %GC in the 100 bp of DNA flanking the repeats was positively associated with relative expandability (Fig. 3). These correlations were particularly high for male transmissions determined by pedigree analysis (*rho* = 0.817, $P < 0.01$, $n = 9$) and directly in sperm (*rho* = 0.9, $P = 0.05$, $n = 5$). Indeed, if the mean relative expandability for the sperm data were used, the *SCA3* data could be included and the correlation was even

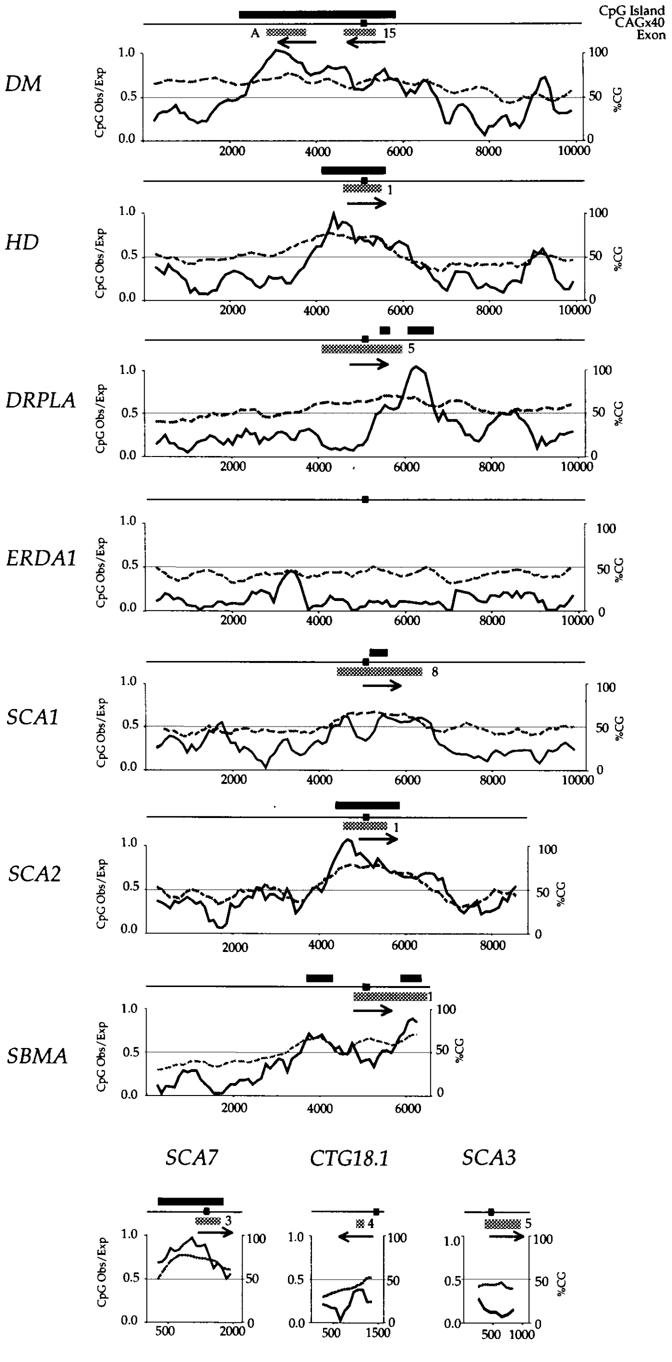

Figure 2. Flanking DNA analyses at the expanded CAG/CTG loci. Shown are diagrammatic representations of the %GC and $\mathrm{CpG}_{\mathrm{obs/exp}}$ plots of the surrounding genomic DNA for each locus up to 5 kb (where available). The %GC is shown as a dashed line, the $\mathrm{CpG}_{\mathrm{obs/exp}}$ as a solid line, and a faint line indicates the position of 50% GC and 0.5 $\mathrm{CpG}_{\mathrm{obs/exp}}$. The 40 × CAG is shown as a small black box, CGIs are shown as black rectangles and the closest exons to the repeat are shown as grey rectangles. Exon numbers are indicated, and an arrow below shows the direction of transcription. Exon 15 of *DMPK* and exon A of *DMAHP* are both shown at the *DM* locus.

more dramatic (*rho* = 0.947, $P < 0.01$, $n = 6$) (data not shown). The correlation between relative expandability during maternal transmission and flanking %GC levels was not as great, but still significant (*rho* = 0.717, $P < 0.025$, $n = 9$). Similar correlations were observed using the mean relative expandability and/or the GC content in the 500 bp of flanking DNA and calculating the expandability as a proportion of total allele length (i.e. length change/progenitor allele) (data not shown).

DISCUSSION

Anecdotal comparisons have previously suggested that the relative stabilities of the expanded triplet repeat loci differ. In this study, we have produced what is, to our knowledge, the first comprehensive quantification of expanded CAG/CTG mutability during germline transmission. This survey involved collating published data regarding the germline transmission of the repeats as determined through pedigree and sperm DNA analyses. To provide a measure of mutability that would be independent of progenitor allele length and, thus, comparable between loci, we have calculated the 'relative length change per expanded progenitor repeat'. This figure essentially reflects the propensity of an expanded allele to undergo further expansion. No correlation between average progenitor allele length and relative expandability was observed, indicating that differences in the relative expandability of the loci did not arise as artefacts of the allele length distributions. Using this method, we have shown that significant differences between these loci do indeed exist. In particular, the *DM* locus appeared to be the most unstable repeat, with the largest length change transmissions observed in pedigree analysis. However, direct sperm DNA analysis suggests that the *SCA7* repeat is actually more unstable than the *DM* repeat during male germline transmission. This phenomenon was not revealed by pedigree analysis, probably due to strong negative selection against the transmission of large repeats in *SCA7* (D.G. Monckton, manuscript in preparation). It is possible that some of the data from the other polyglutamine repeat disorders have been underestimated similarly by pedigree analysis. Indeed, it would appear that data derived from sperm DNA analysis tend to record slightly higher average length changes than those derived from pedigree analysis. However, it remains unclear whether such deviations arise as a result of phenotypic selection or an older average age of sperm donors. Nonetheless, since dramatic deviations from the expected pedigree structures have not been observed for these loci, it is likely that selection has not seriously compromised these estimates.

The measured differences in relative mutability between the loci appear not to be accountable for in terms of repeat length and/or selection effects. Therefore, it seems logical to assume that these differences must be as a result of the influence of *cis*-acting flanking sequences on the mutation process. Intra-locus variations in relative mutability of tandem repeats have previously been associated with subtle flanking sequence polymorphisms (24,26). However, no indications have yet emerged of what the *cis*-acting modifiers of inter-locus variation in humans might be. To this end, we have shown a correlation between the GC content flanking each repeat and their relative mutability. It is of interest that these effects were less pronounced in female transmissions than in male transmission, suggesting that the mutation process is less susceptible to flanking sequence modifiers in females. Data from transgenic mice with expanded CTG repeats have also indicated that this may be a more general phenomenon in mammalian gametogenesis (27).

Whether these correlations indicate a direct cause and effect relationship remains unclear at this point. It is possible that GC

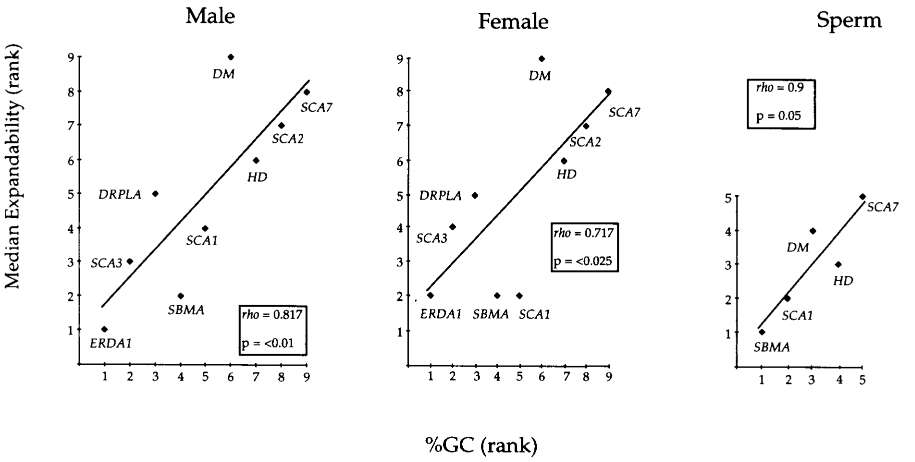

Figure 3. Relative expandability and flanking GC content correlations. Median expandability (Figure 1) at each locus was ranked from one to nine (y-axis), with *DM* the most unstable and *ERDA1* the least, and correlated with the ranked %GC (Table 1) on either flank over 100 bp (x-axis). Median expandabilities in sperm were ranked and similarly correlated with their respective %GC levels. Spearman's coefficient of correlation (*rho*) is shown as is the probability (P) of these associations arising by chance.

content and relative mutabilities of repeat sequences are both responding to some higher order effect of genomic organization, such as location within different isochore subtypes (28). For instance, it is likely that the *DM*, *SCA7*, *SCA2* and *HD* expanded repeats are located in GC-rich H3 isochores (29). Conversely, the *ERDA1* repeat exists in a very large GC-poor region of the genome (40% GC over 92 kb), which is likely to be in an L1 or L2 isochore (28). However, the correlation coefficient is lower when relative mutabilities are compared with the %GC in the 500 bp flanking each repeat as opposed to the 100 bp flanking the repeat, favouring a more localized effect. Nonetheless, closer examination of the 100 bp on each flank revealed no conserved sequence motifs, only large variations in %GC levels. GC content has been shown previously to affect the conformational properties of DNA (30), and it is possible that flanking DNA structures modify repeat metabolism.

It was noted previously that most (11/12) of the expanded triplet repeats then known were close to a CGI, implying that an association with a CGI might be critical for expansion beyond the normal range (31). However, more sequence data and more loci are now available which demonstrate that a number of expanded repeats are not associated with CGIs. These data indicate that proximity to a CGI is not absolutely required in the primary amplification process, although it remains possible that this is still an important factor. Rather, in this study, we have shown that the most unstable expanded CAG/CTG arrays are located within CGIs, whereas the remaining loci are not. Moreover, there are even more subtle correlations between the relative mutabilities of the repeats and the GC content of the flanking DNA.

Whether these associations reflect one and the same or two different mechanisms is difficult to distinguish at this point, since CGIs are by definition regions with high GC content (25). Functional CGIs frequently include the promoter, first exon and start of transcription of an associated gene, and ordinarily are free of methylation at all CpGs (32). Interestingly, although the flanking GC content of the most unstable repeat, *DM*, is not as high as that of some of the other loci, it is located in a very large (~3 kb) unmethylated CGI (33). As neither *SCA7*, *HD* or *SCA2* are X-linked, it is likely that these repeats also occur in hypomethylated stretches of an otherwise densely methylated genome (32). The five other repeats are not located within CGIs and are likely to be surrounded by methylated DNA. Differences in methylation status might be associated with differences in the complement of DNA-binding proteins, altered chromatin state and/or DNA conformation, which in turn may affect the stability of the associated repeat. Mammalian CGIs are known sites for the initiation of transcription and have also been assigned putatively as replication origins (34). Modification of triplet repeat instability has previously been associated with both transcription and replication origin effects in model organisms (35–37).

In summary, these data demonstrate that significant differences in the relative stability of the expanded CAG/CTG repeats exist, independent of allele length, indicating a role for flanking DNA sequences in modifying repeat stability. Moreover, we have revealed striking correlations between relative expandability and flanking GC content and proximity to CGIs, providing the first insights into the nature of global *cis*-acting mutational modifiers at human tandem repeat loci.

MATERIALS AND METHODS

Transmission analyses

For each locus, published references as listed in the relevant sections of the Online Mendelian Inheritance in Man database were examined and data regarding intergenerational transmissions of expanded repeats (progenitor alleles >35 repeats) extracted (OMIM, Center for Medical Genetics, Johns Hopkins University, Baltimore, MD and National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD; http://www.ncbi.nlm.nih.gov/omim). These analyses included all of the transmissions reported, including expansions, contractions and zero length change transmissions. References not currently listed in OMIM are shown in Table 1. The 'relative length change per expanded progenitor repeat' was calculated as length change/(progenitor allele length – 35 repeats). In cases where transmission data did not include progenitor allele size, we used the average progenitor allele length for that locus determined in other studies. The median length change and 95% confidence intervals at each of the loci were calculated using the Wilcoxon signed rank test and MINITAB software. The sex-average figures represent the mean of the male and female median figures. In the data collected for the *DM* locus, transmissions from progenitor alleles of >100 repeats were not included. Transmissions from such large alleles have only been observed at the *DM* locus and they appear to lie outside the linear range, frequently resulting in the transmission of hundreds of extra repeats (38,39). At the *SCA3* locus, six repeats were subtracted from the reported allele lengths so as not to include the six cryptic repeats at the beginning of the array in the expandability calculations.

Sequence analyses

Sequences were obtained from published references and using the Blast facility at NCBI (40) (relevant accession numbers are shown in Table 1). Sequences were first edited to standardize the repeat size at each (40 × CAG) before analysis using the GCG9 program accessed through the HGMP database (http://www.hgmp.mrc.ac.uk/). Each sequence was examined in the 5' to 3' orientation, with the top strand containing the 40 × CAG repeat. Sequences were analysed to determine %GC and $\mathrm{CpG}_{\mathrm{obs/exp}}$ using ‘cpgplot’ with a moving window of 500 bp and a step of 100 bp. The NIX program was used to define the positions of potential coding sequences [G.W. Williams, P.M. Woollard and P. Hingamp, ‘NIX: A nucleotide identification system at the HGMP-RC’ http://www.hgmp.mrc.ac.uk/NIX/] and CGIs, using an algorithm based on the definition by Gardiner-Gardner and Frommer (25). The immediate flanking DNA was defined as that sequence immediately 5' and 3' of the pure $(\mathrm{CAG})_n$ tract, i.e. the six cryptic repeats were included as flanking DNA at the *SCA3* locus. The 100 bp sequences immediately flanking the repeats were also examined for shared sequence motifs and possible alignments using the Gene Jockey program (Biosoft).

ACKNOWLEDGEMENTS

We would like to thank the University of Glasgow Dynamic Mutation Group and Richard Wilson for helpful discussions with this work. D.G.M. is a Lister Institute Research Fellow. This work was also supported by awards from the Medical Research Council (UK) and the Muscular Dystrophy Association (USA).

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.W. and Zoghbi, H.Y. (1993) Expansion of an unstable CAG repeat in spinocerebellar ataxia type 1. *Nature 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., Durr, A., Didierjean, O., Stevanin, G., Agid, Y. and Brice, A. (1996) Cloning of the gene for spinocerebellar ataxia 2 reveals a locus with high sensitivity to expanded CAG/glutamine repeats. *Nature Genet.*, **14**, 285–291.
3. Kawaguchi, Y., Okamoto, T., Taniwaki, M., Aizawa, M., Inoue, M., Katayama, S., Kawakami, H., Nakamura, S., Nishimura, M., Akiguchi, I., Kimura, J., Narumiya, S. and Kakizuka, A. (1994) CAG expansions in a novel gene for Machado–Joseph disease at chromosome 14q32.1. *Nature Genet.*, **8**, 221–227.
4. David, G., Abbas, N., Stevanin, G., Dürr, A., Yvert, G., Cancel, G., Weber, C., Imbert, G., Saudou, F., Antoniou, E., Drabkin, H., Gemmill, R., Giunti, P., Benomar, A., Wood, N., Ruberg, M., Agid, Y., Mandel, J.L. and Brice, A. (1997) Cloning of the SCA7 gene reveals a highly unstable CAG repeat expansion. *Nature Genet.*, **17**, 65–70.
5. Koide, R., Ikeuchi, T., Onodera, O., Tanaka, H., Igarashi, S., Endo, K., Takahashi, H., Kondo, R., Ishikawa, A., Hayashi, T., Saito, M., Tomoda, A., Miike, T., Naito, H., Ikuta, F. and Tsuji, S. (1994) Unstable expansion of CAG repeat in hereditary dentatorubral-pallidoluysian atrophy (DRPLA). *Nature Genet.*, **6**, 9–13.
6. La Spada, A.R., Wilson, E.M., Lubahn, D.B., Harding, A.E. and Fischbeck, K.H. (1991) Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. *Nature*, **352**, 77–79.
7. The Huntington’s Disease Collaborative Research Group (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington’s disease chromosomes. *Cell*, **72**, 971–983.
8. Fu, Y.H., Pizzuti, A., Fenwick, R.G., King, J., Rajnarayan, S., Dunne, P.W., Dubel, J., Nasser, G.A., Ashizawa, T., de Jong, P., Wieringa, B., Korneluk, R., Perryman, M.B., Epstein, H.F. and Caskey, C.T. (1992) An unstable triplet repeat in a gene related to myotonic muscular dystrophy. *Science*, **255**, 1256–1258.
9. Brook, J.D., McCurrach, M.E., Harley, H.G., Buckler, A.J., Church, D., Aburatani, H., Hunter, K., Stanton, V.P., Thirion, J.-P., Hudson, T., Sohn, R., Zemelman, B., Snell, R.G., Rundle, S.A., Crow, S., Davies, J., Shelbourne, P., Buxton, J., Jones, C., Juvonen, V., Johnson, K., Harper, P.S., Shaw, D.J. and Housman, D.E. (1992) 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.
10. Mahadevan, M., Tsilfidis, C., Sabourin, L., Shutler, G., Amemiya, C., Jansen, G., Neville, C., Narang, M., Barcelo, J., O’Hoy, K., Leblond, S., Earle-Macdonald, J., de Jong, P.J., Wieringa, B. and Korneluk, R.G. (1992) Myotonic dystrophy mutation: an unstable CTG repeat in the 3' untranslated region of the gene. *Science*, **255**, 1253–1255.
11. Boucher, C.A., King, S.K., Carey, N., Krahe, R., Winchester, C.L., Rahman, S., Creavin, T., Meghji, P., Bailey, M.E.S., Chartier, F.L., Brown, S.D., Siciliano, M.J. and Johnson, K.J. (1995) A novel homeodomain-encoding gene is associated with a large CpG island interrupted by the myotonic dystrophy unstable $(\mathrm{CTG})_n$ repeat. *Hum. Mol. Genet.*, **4**, 1919–1925.
12. Breschel, T.S., McInnis, M.G., Margolis, R.L., Sirugo, G., Corneliussen, B., Simpson, S.G., McMahon, F.J., MacKinnon, D.F., Xu, J.F., Pleasant, N., Huo, Y., Ashworth, R.G., Grundstrom, C., Grundstrom, T., Kidd, K.K., DePaulo, J.R. and Ross, C.A. (1997) A novel, heritable, expanding CTG repeat in an intron of the SEF2-1 gene on chromosome 18q21.1. *Hum. Mol. Genet.*, **6**, 1855–1863.
13. Ikeuchi, T., Sanpei, K., Takano, H., Sasaki, H., Tashiro, K., Cancel, G., Brice, A., Bird, T.D., Schellenberg, G.D., Pericak-Vance, M.A., Welsh-Bohmer, K.A., Clark, L.N., Wilhelmsen, K. and Tsuji, S. (1998) A novel long and unstable CAG/CTG trinucleotide repeat on chromosome 17q. *Genomics*, **49**, 321–326.
14. Nakamoto, M., Takebayashi, H., Kawaguchi, Y., Narumiya, S., Taniwaki, M., Nakamura, Y., Ishikawa, Y., Akiguchi, I., Kimura, J. and Kakizuka, A. (1997) A CAG/CTG expansion in the normal population. *Nature Genet.*, **17**, 385–386.
15. Harper, P.S., Harley, H.G., Reardon, W. and Shaw, D.J. (1992) Anticipation in myotonic dystrophy: new light on an old problem. *Am. J. Hum. Genet.*, **51**, 10–16.
16. Ashley, C.T. Jr and Warren, S.T. (1995) Trinucleotide repeat expansion and human disease. *Annu. Rev. Genet.*, **29**, 703–728.
17. 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. *Nature Genet.*, **5**, 254–258.
18. Cancel, G., Durr, A., Didierjean, O., Imbert, G., Burk, K., Lezin, A., Belal, S., Benomar, A., Abada-Bendib, M., Vial, C., Guimaraes, J., Chneiweiss, H., Stevanin, G., Yvert, G., Abbas, N., Saudou, F., Lebre, A.S., Yahyaoui, M., Hentati, F., Vernant, J.C., Klockgether, T., Mandel, J.L., Agid, Y. and Brice, A. (1997) Molecular and clinical correlations in spinocerebellar ataxia 2: a study of 32 families. *Hum. Mol. Genet.*, **6**, 709–715.
19. Chong, S.S., Almqvist, E., Telenius, H., LaTray, L., Nichol, K., Bourdelat-Parks, B., Goldberg, Y.P., Haddad, B.R., Richards, F., Sillence, D., Greenberg, C.R., Ives, E., Van den Engh, G., Hughes, M.R. and Hayden, M.R. (1997) Contribution of DNA sequence and CAG size to mutation

frequencies of intermediate alleles for Huntington disease: evidence from single sperm analyses. *Hum. Mol. Genet.*, **6**, 301–309.

20. Zhang, L., Fischbeck, K.H. and Arnheim, N. (1995) CAG repeat length variation in sperm from a patient with Kennedy's disease. *Hum. Mol. Genet.*, **4**, 303–305.
21. Monckton, D.G., Wong, L.-J.C., Ashizawa, T. and Caskey, C.T. (1995) Somatic mosaicism, germline expansions, germline reversions and intergenerational reductions in myotonic dystrophy males: small pool PCR analyses. *Hum. Mol. Genet.*, **4**, 1–8.
22. Leeflang, E.P., Zhang, L., Tavaré, S., Hubert, R., Srinidhi, J., MacDonald, M.E., Myers, R.H., de Young, M., Wexler, N.S., Gusella, J.F. and Arnheim, N. (1995) Single sperm analysis of the trinucleotide repeats in the Huntington's disease gene: quantification of the mutation frequency and spectrum. *Hum. Mol. Genet.*, **4**, 1519–1526.
23. Chong, S.S., McCall, A.E., Cota, J., Subramony, S.H., Orr, H.T., Hughes, M.R. and Zoghbi, H.Y. (1995) Gametic and somatic tissue specific heterogeneity of the expanded SCA1 CAG repeat in spinocerebellar ataxia type 1. *Nature Genet.*, **10**, 344–350.
24. Takiyama, Y., Sakoe, K., Soutome, M., Namekawa, M., Ogawa, T., Nakano, I., Igarashi, S., Oyake, M., Tanaka, H., Tsuji, S. and Nishizawa, M. (1997) Single sperm analysis of the CAG repeats in the gene for Machado–Joseph disease (MJD1): evidence for non-Mendelian transmission of the MJD1 gene and for the effect of the intragenic CGG/GGG polymorphism on the intergenerational instability. *Hum. Mol. Genet.*, **6**, 1063–1068.
25. Gardiner-Gardner, M. and Frommer, M. (1987) CpG islands in vertebrate genomes. *J. Mol. Biol.*, **196**, 261–282.
26. Monckton, D.G., Neumann, R., Guram, T., Fretwell, N., Tamaki, K., MacLeod, A. and Jeffreys, A.J. (1994) Minisatellite mutation rate variation associated with a flanking DNA sequence polymorphism. *Nature Genet.*, **8**, 162–170.
27. Monckton, D.G., Coolbaugh, M.I., Ashizawa, K.T., Siciliano, M.J. and Caskey, C.T. (1997) Hypermutable myotonic dystrophy CTG repeats in transgenic mice. *Nature Genet.*, **15**, 193–196.
28. Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M. and Rodier, F. (1985) The mosaic genome of warm-blooded vertebrates. *Science*, **228**, 953–958.
29. Saccone, S., Caccio, S., Kusuda, J., Andreozzi, L. and Bernardi, G. (1996) Identification of the gene-richest bands in human chromosomes. *Gene*, **174**, 85–94.
30. Gellibolian, R. and Bacolla, A. (1998) Biophysical and structural studies on triplet repeat sequences: duplex triplet repeat structures. In Wells, R.D. and Warren, S.T. (eds), *Genetic Instabilities and Hereditary Neurological Diseases*. Academic Press, London, pp. 561–583.
31. Gourdon, G., Dessen, P., Lia, A.S., Junien, C. and Hofmann-Radvanyi, H. (1997) Intriguing association between disease associated unstable trinucleotide repeat and CpG island. *Ann. Genet.*, **40**, 73–77.
32. Bird, A.P. (1987) CpG islands as gene markers in the vertebrate nucleus. *Trends Genet.*, **3**, 342–347.
33. Shaw, D.J., Chaudhary, S., Rundle, S.A., Crow, S., Brook, J.D., Harper, P.S. and Harley, H.G. (1993) A study of DNA methylation in myotonic dystrophy. *J. Med. Genet.*, **30**, 189–192.
34. Delgado, S., Gomez, M., Bird, A. and Antequera, F. (1998) Initiation of DNA replication at CpG islands in mammalian chromosomes. *EMBO J.*, **17**, 2426–2435.
35. Kang, S., Jaworski, A., Ohshima, K. and Wells, R.D. (1995) Expansion and deletion of CTG repeats from human disease genes are determined by the direction of replication in *Escherichia coli*. *Nature Genet.*, **10**, 213–218.
36. Bowater, R.P., Jaworski, A., Larson, J.E., Parniewski, P. and Wells, R.D. (1997) Transcription increases the deletion frequency of long CTG/CAG triplet repeats from plasmids in *Escherichia coli*. *Nucleic Acids Res.*, **25**, 2861–2868.
37. Miret, J.J., Pessoa-Brandao, L. and Lahue, R.S. (1998) Orientation-dependent and sequence-specific expansions of CTG/CAG trinucleotide repeats in *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA*, **95**, 12438–12443.
38. Redman, J.B., Fenwick, R.G., Fu, Y.-H., Pizzuti, A. and Caskey, C.T. (1993) Relationship between parental trinucleotide GCT repeat length and severity of myotonic dystrophy in offspring. *J. Am. Med. Assoc.*, **269**, 1960–1965.
39. Barceló, J.M., Mahadevan, M.S., Tsilfidis, C., MacKenzie, A.E. and Korneluk, R.G. (1993) Intergenerational stability of the myotonic dystrophy protomutation. *Hum. Mol. Genet.*, **2**, 705–709.
40. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
41. David, G., Durr, A., Stevanin, G., Cancel, G., Abbas, N., Benomar, A., Belal, S., Lebre, A.S., Abada-Bendib, M., Grid, D., Holmberg, M., Yahyaoui, M., Hentati, F., Chkili, T., Agid, Y. and Brice, A. (1998) Molecular and clinical correlations in autosomal dominant cerebellar ataxia with progressive macular dystrophy (SCA7). *Hum. Mol. Genet.*, **7**, 165–170.
42. Gouw, L.G., Castañeda, M.A., McKenna, C.K., Digre, K.B., Pulst, S.M., Perlman, S., Lee, M.S., Gomez, C., Fischbeck, K., Gagnon, D., Storey, E., Bird, T., Jeri, F.R. and Ptácek, L.J. (1998) Analysis of the dynamic mutation in the SCA7 gene shows marked parental effects on CAG repeat transmission. *Hum. Mol. Genet.*, **7**, 525–532.
43. Del-Favero, J., Krols, L., Michalik, A., Theuns, J., Löfgren, A., Goossens, D., Wehnert, A., Van den Bossche, D., Van Zand, K., Backhovens, H., van Regenmorter, N., Martin, J.J. and Van Broeckhoven, C. (1998) Molecular genetic analysis of autosomal dominant cerebellar ataxia with retinal degeneration (ADCA type II) caused by CAG triplet repeat expansion. *Hum. Mol. Genet.*, **7**, 177–186.
44. 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.
45. Telenius, H., Almqvist, E., Kremer, B., Spence, N., Squitieri, F., Nichol, K., Grandell, U., Starr, E., Benjamin, C., Castaldo, I., Calabrese, O., Anvret, M., Goldberg, Y.P. and Hayden, M.R. (1995) Somatic mosaicism in sperm is associated with intergenerational $(CAG)_n$ changes in Huntington disease. *Hum. Mol. Genet.*, **4**, 189–195.
46. Shimohata, T., Ishiguro, H., Makino, K., Takano, H., Tanaka, H., Tsuji, S. and Hirota, K. (1998) Sporadic cases of dentatorubral-pallidoluysian atrophy associated with maternal transmission. *Neurology*, **50**, 282–283.
47. Ikeuchi, T., Onodera, O., Oyake, M., Koide, R., Tanaka, H. and Tsuji, S. (1995) Dentatorubral-pallidoluysian atrophy (DRPLA): close correlation of CAG repeat expansions with the wide spectrum of clinical presentations and prominent anticipation. *Semin. Cell Biol.*, **6**, 37–44.
48. Biancalana, V., Serville, F., Pommier, J., Julien, J., Hanauer, A. and Mandel, J.L. (1992) Moderate instability of the trinucleotide repeat in spino bulbar muscular atrophy. *Hum. Mol. Genet.*, **1**, 255–258.