A Role for Selection in Regulating the Evolutionary Emergence of Disease-Causing and Other Coding CAG Repeats in Humans and Mice

John M. Hancock,† Elizabeth A. Worthey,*‡ and Mauro F. Santibáñez-Koref**

*MRC Clinical Sciences Centre, Imperial College School of Medicine, Hammersmith Hospital, London, England; †Department of Computer Science, Royal Holloway University of London, Egham, Surrey, England; and ‡Leishmania Genome Group, Seattle Biomedical Research Institute, Seattle, Washington

The evolutionary expansion of CAG repeats in human triplet expansion disease genes is intriguing because of their deleterious phenotype. In the past, this expansion has been suggested to reflect a broad genomewide expansion of repeats, which would imply that mutational and evolutionary processes acting on repeats differ between species. Here, we tested this hypothesis by analyzing repeat- and flanking-sequence evolution in 28 repeat-containing genes that had been sequenced in humans and mice and by considering overall lengths and distributions of CAG repeats in the two species. We found no evidence that these repeats were longer in humans than in mice. We also found no evidence for preferential accumulation of CAG repeats in the human genome relative to mice from an analysis of the lengths of repeats identified in sequence databases. We then investigated whether sequence properties, such as base and amino acid composition and base substitution rates, showed any relationship to repeat evolution. We found that repeat-containing genes were enriched in certain amino acids, presumably as the result of selection, but that this did not reflect underlying biases in base composition. We also found that regions near repeats showed higher nonsynonymous substitution rates than the remainder of the gene and lower nonsynonymous rates in genes that contained a repeat in both the human and the mouse. Higher rates of nonsynonymous mutation in the neighborhood of repeats presumably reflect weaker purifying selection acting in these regions of the proteins, while the very low rate of nonsynonymous mutation in proteins containing a CAG repeat in both species presumably reflects a high level of purifying selection. Based on these observations, we propose that the mutational processes giving rise to polyglutamine repeats in human and murine proteins do not differ. Instead, we propose that the evolution of polyglutamine repeats in proteins results from an interplay between mutational processes and selection.

Introduction

Human triplet expansion diseases are predominantly neurological and are caused by instability and expansion of tandem repeats of triplet motifs within or near genes (reviewed in Rubinsztein 1999). The largest class of these diseases results from the expansion of $\underline{\text{C}}$AG repeats within exons. (Throughout this paper, codon repeats that occupy a particular reading frame are designated by underlining the base in the first codon position, e.g., $\underline{\text{C}}$AG. Otherwise, repeats may be considered to be in any frame.) An intriguing feature of these disease-causing repeats is that they have apparently undergone evolutionary expansion. Repeats in these genes are generally absent in rodent homologs, and comparative studies indicate an increase in repeat length during primate evolution, with humans generally having the longest repeats (Rubinsztein et al. 1994; 1995*b*; Djian, Hancock, and Chana 1996).

Two explanations for these observations have been proposed. The first suggests that the evolutionary expansion of these repeats reflects their genomewide expansion along the primate lineage and especially in humans (Rubinsztein et al. 1995*a*). The reality of such lineage-specific, genomewide effects remains uncertain, despite a number of subsequent analyses (reviewed in Amos 1999; Rubinsztein, Amos, and Cooper 1999). This is primarily because of the confounding effect of ascertainment bias (Ellegren, Primmer, and Sheldon 1995), that is, the expectation that repeats isolated in one species will be longer than their homologs in other species as they have been isolated because of their polymorphic nature. Long repeats are more polymorphic than short repeats. Ascertainment bias confounds even the relatively well studied comparison between humans and chimpanzees, while evidence for such differences between humans and other primates is lacking, and indeed there is some evidence to the contrary (e.g., Morin et al. 1998). There is also evidence for very long CAG repeats in mice (King et al. 1998). A number of explanations have been suggested for the human-chimpanzee difference (Amos 1999; Rubinsztein, Amos, and Cooper 1999), but these rely on characteristics of human and chimpanzee evolutionary history and therefore cannot provide an explanation for changes in repeat length over long periods of evolution.

The second possible explanation for the evolutionary expansion of $\underline{\text{C}}$AG repeats in these genes is that forces or processes that are specific to individual genes and/or genomic locations act on particular genes in particular evolutionary lineages to give rise to locus- and lineage-specific expansions. One prominent candidate for such an influence is local base (and nucleotide motif) composition. Different isochores in mammalian genomes have different GC compositions, and genes within these regions show correlated base compositions, notably at third codon positions (Mouchiroud, Gautier, and Bernardi 1995). Thus, genes within GC-rich isochores will tend to accumulate concentrations of codons with G and C at their third positions, which might act as seeds for replication slippage and predispose genes to

Key words: CAG repeats, triplet expansion diseases, simple sequences, natural selection.

Address for correspondence and reprints: John M. Hancock, Department of Computer Science, Royal Holloway University of London, Egham, Surrey TW20 0EX, United Kingdom. E-mail: j.hancock@dcs.rhul.ac.uk.

Mol. Biol. Evol. 18(6):1014–1023. 2001

© 2001 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

accumulating codon repeats. In the extreme, such biases could even bias amino acid compositions of proteins, again predisposing genes to seeding of codon repeats (Nakachi et al. 1997; Nishizawa and Nishizawa 1998; Brock, Anderson, and Monckton 1999). Brock, Anderson, and Monckton (1999) have even suggested that local base composition affects the frequency of indel mutations at CAG repeats. Another possibility is that of the effects of local mutation rate. Kruglyak et al. (1998) have suggested that the equilibrium length of microsatellites is a consequence of the balance between the rates of point and slippage mutation. Incorporation of point mutations into repeats reduces their rate of length change during evolution (Albà, Santibáñez-Koref, and Hancock 1999*a*). If either or both of these parameters varied across a genome, this could affect the accumulation of tandem repeats. Finally, Djian, Hancock, and Chana (1996) have suggested that codon repeats in disease genes are flanked by regions with a relatively high frequency of acceptance of point mutations. Mutational instability of regions immediately flanking CA microsatellites has also been suggested by Brohede and Ellegren (1999). High rates of sequence change could reflect a relatively low level of purifying selection in the vicinity of repeats. Selective forces could differ between genes and subregions of genes, depending on the phenotypic consequences of mutations in these different locations. These differences could affect the probability of tandem repeats arising, and, in particular, expanding, during evolution (Nishizawa, Nishizawa, and Kim 1999). The recent demonstration for *Saccharomyces cerevisiae* that transcription factors and protein kinases are significantly overrepresented among proteins that contain polyglutamine repeats (Albà, Santibáñez-Koref, and Hancock 1999*b*) also indicates a role for selective constraints in the evolution of these structures, although their functional significance remains unclear (Schmid and Tautz 1999).

Here, we addressed the question of the forces giving rise to the evolutionary expansion of CAG repeats in triplet expansion disease and other genes by comparing the lengths of CAG repeats in humans and mice and by considering the base and codon compositions and rates of synonymous and nonsynonymous substitution in CAG repeat-containing genes. We found no evidence of a preferential accumulation of CAG repeats in the human genome relative to the mouse genome or of differences in the nature of the selection acting on genic positioning of CAG repeats in the two species. When we considered pairs of proteins that contained a CAG repeat in one species but not the other, we found no differences in the properties of surrounding sequences. However, we did find an overrepresentation, relative to the average amino acid usage in humans and mice, of the amino acids proline, glutamine, histidine, and serine, which may have given rise to biases in the gene sequences and predisposed them to accumulating repeats. We also observed locally high levels of nonsynonymous base substitution in the neighborhood of repeats in genes containing a repeat in only one species, but low levels in genes in which repeats were conserved between humans and mice. We combine these observations to propose a hypothesis to explain the evolution of these repeats.

Materials and Methods

Database Screening and Analysis

Genes containing repeats of five or more CAG codons in humans (*Homo sapiens*), mice (*Mus musculus*), or both were identified from a data set described previously (Albà, Santibáñez-Koref, and Hancock 1999*a*). This data set was compiled by screening the human and mouse subsets of GenBank for proteins with tracts of six or more glutamines using BLASTP (Altschul et al. 1990) and eliminating redundancy in the data by running FASTA (Pearson and Lipman 1988). Database entries were obtained using ENTREZ at the National Center for Biotechnology Information, Bethesda, Md. (http://www.ncbi.nlm.nih.gov/entrez/). Sequences with 95% identity were considered redundant, and only one representative was used in subsequent analysis. Discrepancies in the lengths of polyglutamine tracts in nearly identical sequences were resolved by taking the sequence with the longest tract. BLASTP was then used to identify homologous sequences from the other species, and sequence similarity was confirmed using the GCG program PILEUP (Genetics Computer Group 1997). Members of this data set that contained CAG repeats of length 5 or greater in at least one species were then identified and classified into three groups: genes containing a CAG repeat in both humans and mice (group B); genes containing repeats in humans but not in mice (group H); and genes containing a CAG repeat in mice but not in humans (group M).

For comparative analysis of database sequences containing CAG repeats of length 7 or more, the GenBank and EMBL DNA databases, including EST and STS subgroups, were analyzed using routines from the GCG package, version 9.1 (Genetics Computer Group 1997), unless otherwise noted. The databases were searched using the pattern recognition routine FINDPATTERNS. Entries showing >95% identity to one another upon multiple sequence alignments using PILEUP (Genetics Computer Group 1997), CLUSTAL W (Thompson, Higgins, and Gibson 1994), version 1.7, and FASTA (Pearson and Lipman 1988) were considered to represent the same sequence and grouped together. This allowed for sequencing errors without grouping members of gene families together as single loci. The sequence with the longest array was again taken as the representative from each of these groups. Database entries were again obtained using ENTREZ. The genic locations of repeats were identified using sequence annotations where these were available.

Sequence Analysis Methods

Tandem codon arrays of length ≥5 were identified using ARRAYFINDER (Hancock et al. 1999). A modified version of ARRAYFINDER (PROTARRAY) allowed identification of all amino acid tandem repeats of this length. cDNA codon frequencies were calculated using the GCG program CODONFREQUENCY (Genetics Computer Group 1997). These frequencies were used to calculate overall and third-codon-position base compositions using a commercially available spreadsheet, which was also used to carry out most statistical tests. Other statistical tests were carried out using the SPSS package and the VassarStats web server (http://faculty.vassar.edu/~lowry/VassarStats.html). Significance thresholds were subjected to Bonferroni adjustment to take into account multiple testing. Significance values quoted in the text are also Bonferroni-adjusted. Expected amino acid frequencies in cDNAs were calculated on the basis of overall codon frequency tables for mice and humans obtained from the CUTG database server (Nakamura, Gojobori, and Ikemura 2000) at http://www.kazusa.or.jp/codon/. To calculate synonymous and nonsynonymous DNA sequence divergences (K_s and K_a), sequence pairs were aligned using the LaserGene program MEGALIGN (DNASTAR, Madison, Wis.). Alignments were calculated by translating cDNAs into protein sequences and using the method of Hein (1990), which coped better with sequences of unequal length than the Clustal algorithm (Higgins and Sharp 1989) as implemented in MEGALIGN. K_s and K_a for sequence pairs were calculated using MEGA, version 1.01 (Kumar, Tamura, and Nei 1993) using the Jukes-Cantor correction for saturation (Jukes and Cantor 1969). We excluded all repetitive regions from the analysis. Regions to be excluded were initially identified by length difference between species (i.e., presence of an indel in the alignment). The limits of the repeat region were then defined by extending the repeat as far as the last codon adjacent to the repeat that was identical in two out of three positions to the tandemly repeated codon in either species. This excluded not only CAG repeats, but also all other length-varying codon repeats.

Results

Repeat Evolution

We identified 28 genes for which complete cDNA sequences were available for both mice and humans and which contained a $(\mathrm{CAG})_{\geq 5}$ array in at least one species (table 1). Of these genes, 10 contained a CAG array in both species (B genes), 10 (of which 5 were human triplet expansion disease genes) contained a CAG array in the human sequence only (H genes), and 8 contained a CAG array in the mouse sequence only (M genes) (table 1). Thirty-one CAG arrays were identified in 20 human cDNAs, and 31 were identified in 18 mouse cDNAs. Mean CAG repeat lengths were 8.4 for humans and 8.0 for mice. The length distributions were not significantly different ($P = 0.73$, two-tailed Mann-Whitney U test). Group M genes might be expected to reveal any bias in CAG repeat length between humans and mice, as they contain repeats in both species, but no significant difference was detected in these genes (Wilcoxon signed-ranks test, $P > 0.05$, $N = 14$, two-tailed test). Thus, we found no evidence of a difference in CAG repeat length between humans and mice in this data set.

We also screened these sequences for amino acid repeats in the conceptual translation, as amino acid repeats are frequently encoded by mixtures of synonymous codons (Albà, Santibáñez-Koref, and Hancock 1999*a*, 1999*b*) (table 1). Thirty-seven of 81 amino acid repeats of length ≥5 in human proteins were of glutamine, compared with 47 of 82 repeats in mouse proteins. Mean lengths for these repeats were 12.7 for humans and 10.6 for mice (difference not significant, $P = 0.50$, two-tailed Mann-Whitney U test). Within group B, glutamine repeats were significantly longer in humans than in mice (Wilcoxon signed-ranks test, $P < 0.05$, $N = 17$, two-tailed test). The most common other classes of repeats were those of proline (12 in humans, 10 in mice), glycine (9 in humans, 7 in mice), and glutamic acid (7 in humans, 5 in mice). The higher proportion of glutamine repeats with respect to others in the mouse proteins was not significant ($P > 0.05$, chi-square, df = 1). We therefore found the relative tendencies for proteins to accumulate Gln versus other amino acid repeats to be similar in mice and humans. We also found that Gln repeats accumulating in human proteins tended to be longer than those in mouse proteins in group B. This tendency was not observed for the other gene groups.

To further investigate whether the lengths of human and mouse CAG repeats differed, we screened databases for tandem CAG repeats of length >7 in the two species. We identified all repeats, irrespective of their locations within genes, and did not restrict our search to pairs of homologous sequences. Mean lengths (in base pairs) for these repeats were 29.06 (median 27, $N = 205$) for humans and 36.05 (median 33, $N = 63$) for mice. The length distributions were significantly different ($P < 0.001$, Mann-Whitney U test), with mice tending to have longer CAG repeats than humans. We therefore found no bias toward longer CAG repeats in humans versus mice at the whole-genome level and, indeed, found evidence of the opposite bias.

There is no a priori reason to expect tandem repeats of CAG to lie in any particular reading frame of an exon unless selection has constrained the reading frames in which these repeats have been able to expand. Frame specificity of this kind has been reported previously (Stallings 1994). To test for any global difference in this pattern (and therefore in the selection causing it) between humans and mice, we investigated the locations of the identified repeats that lay within adequately annotated database sequences (table 2). CAG repeats were preferentially found in the reading frame encoding glutamine (reading frame 1 in table 2) in both humans and mice ($P < 0.0001$ for mice, humans, and overall; chi-square against an even distribution in all six reading frames, df = 5). There was no significant difference in repeat distribution between species (chi-square test for inhomogeneity in the 2 × 9 contingency table; $P > 0.05$; df = 8). Thus, there appear to be no strong differences in the selective forces acting on the locations of CAG repeats in the human and mouse genomes.

The results described in this section indicate no significantly greater length of CAG repeats in the human genome with respect to that of the mouse or in human

Table 1
Long Amino Acid and Codon Repeats in Gene Pairs Analyzed in this Study

		Human			Mouse		
Group	Gene Name[a]	Accession No.	CAG/CAA Repeats[b]	Glutamine Repeats[c]	Accession No.	CAG/CAA Repeats	Glutamine Repeats
B	ATBF-1	L32832	7, 5, 5, (10)	19, 8, 5, 5, 6	D26046	11, 6, 7, 9	13, 6, 18, 9, 9
	CAGH45/MOPA-1	U80742	5, 5, 7, 6, 7	26, 6, 26, 7, 5	AF071310	6, 5, 5, 5, 5	33, 7, 27, 5, 5
	CBFA1	AF001444	6, 6	23	AF010284	6, 8	29
	DB1/VEZF1	D28118	7, (6)	13	AF104410	6	15
	KIAA0334/CLOCK	AB002332	6	6, 5	AF000998	8	9, 8, 5
	N-OCT-3	L37868	6	21	M88300	11	23
	PQPROT/TDAG51	Z50194	6	14	U44088	5	8
	SATB1	M97287	7	15	U05252	6	15
	TBP (B)	M55654	9, 18	38	U63933	5	15
	TGFβ/VGR-1	M60315	6	6	J04566	10	10
H	AR(D)	M20132	20, 6	21, 6, 5	S56585		8
	DRPLA (D)	D38529	10	14	D87744		
	HD (D)	L12392	21	23	L28827		7
	SCA1 (D)	X79204	12, 15	12, 15	X83542		
	SCA2 (D)	U70323	8, 8	22	AF041472		
	ASH1	L08424	14	14	M95603		
	IRE1	AF059198	6	6	AF071777		
	IRS-1	S85963	6	6	L24563		6
	TIS11d/TIS11	U07802	7	7	M58564		
	TRAM-1/p/CIP	AF016031	6, 9	5, 29	AF000581		23
M.	CHGA	J03915			M64278	11	11
	GCRAR/GCRR	X03225			X04435	8	8
	HBF-1/HFHBF-1	X74142			U36760	6	6
	HOXA10	AF040714			L08757	6	6
	IVL	M13903			L28819	18, 6	5, 19, 6, 5
	TAp63β	AF075432		6	AF075435	6	6
	SRY	L10101			U70653	12, 17, 10, 6, 5, 15	12, 5, 17, 5, 5, 10, 6, 5, 5, 5, 15
	TOB	D38305		5, 6	D78382	7, 6, 9	7, 6, 10

[a] Names are according to EMBL database entries. Where names differ between species, the human name is given first. Groups to which the different genes were allocated are given on the left. Genes belonging to the subgroup D are indicated by a D in parentheses. Groups were defined as follows: group B contains CAG repeats of length ≥5 in both species; group H contains CAG repeats of length ≥5 in the human gene only; subgroup D contains genes associated with human triplet expansion disease; and group M contains CAG repeats of length ≥5 in the mouse gene only.

[b] Tandem repeats of ≥5 CAG or CAA codons; identified using ARRAYFINDER (Hancock et al. 1999). CAA codon repeats are given in parentheses.

[c] Tandem repeats of ≥5 glutamines; identified using modified version of ARRAYFINDER (Hancock et al. 1999).

Table 2
Frequencies of CAG Repeats ($n \geq 7$) in Different Reading Frames and Genic Locations in Annotated Human and Mouse Sequences

Location[a]	Human[b]	Mouse[b]	Total[b]
1 (CAG = Gln)	36	13	49
2 (AGC = Ser)........	11	1	12
3 (GCA = Ala)........	1	0	1
−1 (CTG = Leu)......	14	1	15
−2 (GCT = Ala)......	1	0	1
−3 (TGC = Cys)......	4	0	4
Intron................	3	0	3
5′ UTR	9	0	9
3′ UTR	6	0	6
Total.................	85	15	100

[a] Numbers 1 through 3 and −1 through −3 represent the reading frame in which the repeat was found. The corresponding repeated codon and amino acid are given in parentheses. Other noncoding locations are as given.

[b] Numbers of repeats found.

Table 3
Overall and Third-Codon-Position Base Compositions for Gene Pairs

	Human		Mouse	
Gene Group	GC	GC3	GC	GC3
B	0.559	0.625	0.551	0.625
H	0.597	0.675	0.594	0.650
M..........	0.563	0.662	0.565	0.665
All	0.574	0.653	**0.570**	0.645
Expected	0.529	0.601	0.528	0.597

Note.—Gene groups are as in table 1. GC = overall proportion of G + C. GC3 = proportion of G + C at third codon positions. Values significantly higher than the expected value after correction for multiple testing are in bold.

proteins, and, indeed, the opposite appears to be the case. We did, however, observe a significant tendency for glutamine repeats to be longer in human group B proteins than in the homologous mouse proteins.

Base, Codon, and Amino Acid Composition

As base composition has been proposed to be an important factor in driving CAG repeat evolution (Brock, Anderson, and Monckton 1999), we attempted to identify common sequence properties of genes containing disease-causing CAG repeats and consistent changes in homologs containing repeats relative to homologs not containing repeats by analyzing the base compositions of the cDNA sequences for the 28 gene pairs. For both mouse and human homologs and for all gene groups, G+C compositions were on average higher than expected compositions calculated from the CUTG table of codon frequencies (table 3). The overall mean G+C composition (i.e., for groups B, M, and H pooled) deviated significantly from expectation in mice and humans ($P < 0.05$; two-tailed *t*-test). Third-codon-position base compositions were also higher than expected for all groups, but the pooled difference did not approach significance. Interspecies differences in base composition were not statistically significant. Thus, we found a generally high G+C content in the set of genes in both species, even when the gene did not contain a repeat.

High GC compositions could result from mutational bias at third codon positions, for example, due to the isochore location of the gene in question, or they could reflect the amino acid composition of the encoded proteins (Nakachi et al. 1997; Nishizawa and Nishizawa 1998). To test for a relationship between base composition and amino acid composition, we first tested for significant differences in amino acid composition from expected compositions (based on overall species codon frequencies) in our set of proteins. We did this by calculating chi-square values for the pooled amino acid compositions of groups H, M, and B. As we could not expect these goodness-of-fit values to follow the chi-square distribution a priori because of possible inhomogeneity in the set of all proteins, significance (i.e., the probability of randomly drawing a group of 8 or 10 proteins with the calculated goodness-of-fit value or lower from the set of proteins encoded by the human or mouse genome) was estimated by extracting a set of 18,554 proteins from the CUTG codon usage database with sizes of between 205 and 3,727 amino acids (the size range of our sample of repeat-containing proteins). These proteins were then grouped randomly into groups of 8 or 10, and goodness-of-fit values for amino acid composition were calculated for each group. A total of 185,470 groups of size 8 and 185,450 groups of size 10 were analyzed. Values corresponding to appropriate Bonferroni-adjusted ($n = 3$) significance levels were estimated. Groups B and H showed significant deviation from average amino acid composition ($P < 0.01$), whereas group M did not ($P > 0.05$). The scores achieved by the group M proteins in the two species would only have achieved significance for a group of size 30 or larger.

As these analyses indicated significantly biased amino acid compositions, at least for groups B and H, we then calculated the relative representations of amino acids within the 28 proteins, again calculating expectations based on species codon frequencies (table 4). Significances of the observed/expected (O/E) values so calculated were estimated using the same set of sequences as above, calculating O/E values for the same numbers of random groups of 8 or 10 proteins. Confidence levels were estimated for each amino acid separately after adjusting for multiple tests. In both human and mouse data sets, four amino acids (Gln, Pro, His, and Ser) showed a significant overall excess ($P < 0.05$) and showed an excess in all three groups.

Finally, we investigated whether the observed base compositions of these genes could be explained solely on the basis of their amino acid compositions and average genomic codon usage or whether there was an excess of GC-richness that might be due to codon usage bias. This was done by calculating expected base compositions for proteins given their amino acid compositions and the CUTG synonymous codon usages (table 5). Amino acid composition and global genomic codon usage alone could account for the base compositions of these genes. We conclude that the biased base compositions of these genes are due to their unusual amino

Table 4
Over- and Underrepresentation of Amino Acids in the Different Protein Classes

Amino Acid	B		H		M		All	
	Human	Mouse	Human	Mouse	Human	Mouse	Human	Mouse
Ala	0.96	0.97	**1.29**	**1.31**	0.86	0.82	**1.05**	**1.03**
Arg.	0.91	0.80	0.93	0.98	0.81	0.71	0.89	0.84
Asn	0.87	0.98	0.89	0.82	**1.14**	0.88	0.96	0.89
Asp	0.68	0.76	0.74	0.71	0.77	0.89	0.73	0.79
Cys.	0.67	0.59	0.70	0.68	0.73	0.59	0.70	0.65
Gln.	**2.49**	**2.39**	**1.53**	**1.30**	**1.88**	**2.95**	**1.97**	**2.17**
Glu.	0.70	0.79	0.76	0.76	**1.26**	**1.05**	0.88	0.83
Gly.	0.94	0.89	**1.13**	**1.17**	**1.09**	0.99	**1.05**	**1.02**
His	**1.41**	**1.48**	**1.24**	**1.34**	**1.11**	**1.83**	**1.26**	**1.55**
lle.	0.70	0.74	0.58	0.54	0.59	0.57	0.63	0.60
Leu.	0.86	0.91	0.87	0.88	0.94	0.82	0.89	0.87
Lys.	0.81	0.86	0.70	0.67	**1.06**	**1.00**	0.84	0.83
Met	**1.06**	**1.06**	0.96	0.96	**1.16**	0.99	**1.05**	1.00
Phe.	0.73	0.77	0.79	0.78	0.69	0.85	0.74	0.80
Pro	**1.53**	**1.47**	**1.66**	**1.70**	**1.42**	**1.37**	**1.54**	**1.54**
Ser	**1.18**	**1.11**	**1.48**	**1.48**	**1.15**	**1.06**	**1.28**	**1.27**
Thr.	**1.13**	0.98	0.86	0.86	0.72	0.77	0.92	0.88
Trp	0.67	0.61	0.41	0.47	0.64	0.56	0.57	0.53
Tyr.	0.76	0.83	0.85	0.86	0.91	0.82	0.84	0.84
Val	0.80	0.84	0.76	0.79	0.64	0.55	0.74	0.74
Stop	0.77	0.80	0.57	0.64	**1.06**	0.97	0.78	0.76

Note.—Amino acids are represented by their three-letter codons. Values presented are observed/expected ratios for each amino acid (including stop codons) based on the mean amino acid composition averaged over all members of the group. Expected values are calculated from the overall codon usage for a given species. Values shown in bold are greater than 1.000. Underlined values are significantly greater than 1 ($P < 0.05$ after correction for multiple tests) based on the simulations described in *Materials and Methods*.

acid contents rather than any bias in base composition at synonymous codon sites.

Substitution Rate

The accumulation of CAG repeats in genes might be related to the accumulation of base substitutions in the gene for two reasons. First, purifying selection could constrain the accumulation of repeats such that proteins or protein regions under higher levels of purifying selection would accumulate repeats more slowly than regions under weaker purifying selection, if at all. Second, Kruglyak et al. (1998) have suggested that regions undergoing a relatively higher rate of mutation should accumulate repeats more slowly because repeats in such regions are more likely to incorporate interrupting bases. To a first approximation, K_s for a pair of sequences can be taken as an estimator of the mutation rate, while K_a can act as an estimator of the strength of selection acting on individual genes, although the two values tend to be correlated (Graur 1985; Ticher and Graur 1989). Mean whole-gene (excluding repeat) K_s and K_a values for each group are presented in table 6. Mean values of both were lower for genes of group B than for those of groups H and M. The differences between group B and the pooled groups H and M were significant for K_a ($P < 0.001$; Mann-Whitney U test) but not for K_s ($P > 0.05$).

Table 5
Relationship Between Amino Acid and Base Compositional Bias

Group	GC		GC3	
	Expected	Observed	Expected	Observed
Human				
B	0.558	0.559	0.617	0.625
H	0.596	0.597	0.670	0.675
M	0.562	0.563	0.659	0.662
All	0.573	0.574	0.648	0.653
Mouse				
B	0.550	0.551	0.621	0.625
H	0.594	0.594	0.649	0.650
M	0.559	0.565	0.652	0.665
All	0.568	0.570	0.640	0.645

Note.—Overall G + C composition of gene; GC3 = G + C composition of third codon positions. Expected values are predicted based on protein amino acid content and genomic codon usage. Observed values are measured from the data set.

Table 6
Mean K_a and K_s Values (± SD) for Gene Groups

Group	K_s	K_a
B.	0.423 ± 0.148	0.016 ± 0.012
H.	0.586 ± 0.293	0.084 ± 0.098
M	0.564 ± 0.318	0.148 ± 0.196
All	0.521 ± 0.260	0.078 ± 0.126

To investigate whether repeats appear in regions of high mutation rate or low selection relative to the remainder of the protein in which they are located (Djian, Hancock, and Chana 1996), K_s and K_a values were also calculated for regions of arbitrary length 33 codons upstream and downstream of the repeat (table 7). The positions of exon/intron boundaries were not taken into account in this analysis, as they are not known for many of the cDNAs analyzed. However, regions analyzed were truncated at the N- or the C-terminal end of the encoded protein where applicable, or if they overlapped with another tandem repeat (defined as in *Materials and*

Table 7
Synonymous and Nonsynonymous Divergences for Regions Flanking CAG Repeats

Gene	K_s (Near)[a]	K_s (Dist)[b]	K_a (Near)[c]	K_a (Dist)[d]
ATBF-1	**0.783**	0.423	0.007	**0.024**
CAGH45	0.162	**0.411**	**0.027**	0.022
CBFA1	0.172	**0.225**	0.000	**0.005**
DB1	**0.525**	0.215	0.000	**0.007**
KIAA0334	**0.803**	0.438	**0.034**	0.016
N-OCT-3	**0.403**	0.245	**0.010**	0.000
PQPROT	0.613	**0.698**	0.013	**0.045**
SATB1	**0.868**	0.428	0.007	**0.008**
TBP	**0.949**	0.510	**0.043**	0.000
TGFβ	0.442	**0.522**	**0.059**	0.031
N	6	4	5	5
AR	0.244	**0.548**	**0.107**	0.080
DRPLA	**0.479**	0.452	**0.106**	0.030
HD	0.421	**0.621**	0.000	**0.046**
SCA1	0.366	**0.578**	**0.073**	0.056
SCA2	0.268	**0.394**	**0.145**	0.049
ASH1	0.286	**0.359**	**0.054**	0.006
IRE1	**1.690**	1.345	**1.162**	0.354
IRS-1	**0.640**	0.578	**0.069**	0.045
TIS11d	**0.437**	0.408	**0.061**	0.038
TRAM-1	0.645	**0.723**	**0.150**	0.087
CHGA	**1.396**	0.707	0.164	**0.201**
GCRAR	**0.410**	0.372	**0.071**	0.046
HBF-1	**0.679**	0.160	**0.160**	0.030
HOXA10	**0.468**	0.409	0.050	**0.119**
IVL	0.777	**1.267**	**0.682**	0.591
SRY	**0.792**	0.678	**0.478**	0.218
TAp63β	**0.552**	0.473	**0.013**	0.007
TOB	0.289	**0.361**	**0.071**	0.013
N	10	8	**15**	**3**

NOTE.—Gene names are as in table 1. N = Number of occasions on which the value in the Near column is greater than that in the Dist column, or vice versa, for either K_s or K_a. Counts are shown for Group B and for Group H + M separately. Ratios that deviate significantly from 1:1 ($P < 0.05$) by the sign test are shown in bold.

[a] K_s for a region 33 amino acids either side of the longest CAG repeat.

[b] K_s for the entire sequence excluding the repeat and flanking region. The larger of K_s (All) and K_s (Near) is given in bold.

[c] K_a for a region 33 amino acids either side of the longest CAG repeat.

[d] K_a for the entire sequence excluding the repeat and flanking region.

Methods). The pooled group H and M genes had a significant tendency to show lower K_a values near the repeat. This was not so for group B genes or for K_s in any of the gene groups or overall. This indicates that selection is weaker in the vicinity of repeats in group H and M genes, while this is not the case in group B genes. It also indicates that mutation rates do not differ between the vicinity of repeats and more distant parts of genes.

Substitution rates could be affected by the GC-richness of the sequences, as sequences under pressure to adopt an extreme base composition are unable to accept many substitutions. However, we observed no significant correlations between K_a or K_s and overall or third-codon-position base composition (see also Matassi, Sharp, and Gautier 1999).

If a low K_a value is indicative of relatively strong selection acting on a protein, this might also influence the rate of change of the lengths of repeat regions. We therefore investigated the relationship between K_a and the difference in the length of the longest CAG repeat present in each gene, irrespective of the species in which it was found. K_a correlated positively and significantly with this difference ($r = 0.420$, $P < 0.05$).

In summary, these results indicate an association of new repeats with regions of high K_a (corresponding to regions of low purifying selection) and no association with regions of high K_s (corresponding to a high local mutation rate).

Discussion

We looked for evidence that would support the involvement of various forces in the evolutionary expansion of CAG repeats in human (and murine) genes. We first investigated the possibility of a general accumulation of CAG repeats in the human genome but not in other lineages. We found no evidence for preferential accumulation or expansion of CAG repeats in the human genome relative to that of the mouse by comparing either the numbers of genes in the public databases containing CAG repeats in either species, the lengths of the CAG repeats they contain, or the overall length distributions of anonymous CAG repeats in the databases. The latter analysis indicated longer CAG repeats in the mouse than in the human genome. We found no evidence of any difference in the distribution of CAG repeats within coding regions between the species. While these analyses were subject to biases because of numerous screens for long CAG repeats associated with disease (Riggins et al. 1992; Li et al. 1993; Abbott and Chambers 1994; Jiang et al. 1995; Aoki et al. 1996; Chambers and Abbott 1996; Neri et al. 1996; Bulle et al. 1997; Kim et al. 1997; Margolis et al. 1997; Reddy et al. 1997; Albanese et al. 1998; Pawlak et al. 1998; Zuhlke et al. 1999), given the emphasis that has been placed on searches for human sequences of this type, it is unlikely that the databases are more biased toward long repeats in mice.

Our data also do not support the suggestion that local base composition has driven the accumulation of repeats within the 28 pairs of homologous repeat-containing genes we considered (Jurka and Pethiyagoda 1995; Nakachi et al. 1997; Nishizawa and Nishizawa 1998; Brock, Anderson, and Monckton 1999). Although we found higher GC and GC3 contents than expected for all of the gene groups studied here, this reflected solely the biased amino acid compositions of the gene products and was not the result of any preferential use of synonymous codons with GC-rich third positions, as would be expected if mutation toward a biased base composition were the force driving the observed biases. We also did not find any difference in base composition between genes containing repeats and genes not containing repeats, which would be expected if changes in base composition drove repeat evolution.

Finally, we found no relationship between mutation rate, as indicated by the synonymous substitution rate, and the emergence of repeats during evolution. This is not consistent with a model whereby repeat evolution in a genomic locality reflects the balance between point and slippage mutation rates there (Kruglak et al. 1998). However, there is evidence that substitution rates in regions flanking CA microsatellites correlate inversely with repeat length in a larger data set (unpublished data). It is therefore possible that effects of this kind also contribute to the evolution of CAG repeats in genes but that these effects are relatively weak in this data set and/or could not be detected here because of the data set's relatively small size and the correlation between K_a and K_s.

We found three strong patterns in our data set: overrepresentations of certain amino acids, differences in the nonsynonymous substitution rates observed in group B genes compared with group H and M genes, and elevated nonsynonymous substitution rates in the vicinity of repeats in group H and M genes. At the level of amino acid composition, we observed significant overrepresentation of four amino acids, Gln, Pro, Ser, and His, in all genes studied. Along with Gln repeats, we also observed numerous Pro repeats in these proteins. It is likely that the biased amino acid compositions of these genes reflect in some way functional selection on these genes. As these amino acid composition biases are similar in human and mouse proteins, this selection must have taken place before the divergence of the two lineages, one of the most ancient eutherian divergences. The shared overrepresentation of these amino acids between species also indicates that changes in amino acid bias have not driven repeat accumulation. However, the biased amino acid compositions of repeat-containing proteins indicate that such bias might provide a breeding ground for new repeats because new repeats contain an unusual concentration of Gln codons and related codons such as CCG (Pro). The preference for polyglutamine repeats to occur in proteins with these amino acid composition biases could therefore reflect either selection favoring polyglutamine repeats in these proteins as part of a selection for a high Gln content, preferential seeding of $\underline{C}$AG repeats in genes with high concentrations of Gln and high GC-content, or both.

We also found a significant difference in overall K_a (but not K_s) between group B proteins and other proteins and a significant bias toward higher K_a (but not K_s) near the Gln repeat in group H+M but not group B proteins. The K_a values for regions flanking repeats in group H+M genes were twice the average for human-mouse sequence pairs calculated by Makalowski and Boguski (1998), 0.201 compared with 0.090, consistent with our suggestion of high rates of sequence change near disease-causing repeats (Djian, Hancock, and Chana 1996), although this difference was not significant (Mann-Whitney *U* test). These observations indicate that there have been considerably larger differences in strength of selection than in mutation rate in these proteins. If a high K_a value indicates a low level of purifying selection, polyglutamine repeats in proteins in groups H and M could have evolved as effectively neutral structures in a low-purifying-selection environment. Repeats in the group B genes, on the other hand, may have been conserved in a high-purifying-selection environment. The significant correlation between K_a and CAG length difference between species is consistent with this.

The stronger purifying selection acting on the polyglutamine repeats in group B proteins is also consistent with the observation of a significant difference in the lengths of polyglutamine repeats of humans and mice in these genes: there may be differences in the strength or type of selection acting on these repeats between the two species. This, in turn, may reflect in some way the functions of these structures in the two species. However, this difference in repeat length appears to be a special property of genes that have a repeat in both species, as lengths of $\underline{C}$AG repeats did not show any evidence of significant difference between species overall. This difference would therefore not appear to be relevant to neutrally evolving repeats, such as those found in the human disease genes.

Whether or not polyglutamine repeats in proteins affect function remains unclear. Sequence analysis has not provided clear evidence for their functional importance (Treier, Pfeifle, and Tautz 1989; Green and Wang 1994; Karlin and Burge 1996; Michalakis and Veuille 1996; Tautz and Nigro 1998; Schmid and Tautz 1999), but biochemical studies have indicated effects on protein-protein interactions (Kazemi-Esfarjani, Trifiro, and Pinsky 1995; Lanz et al. 1995; Pinto and Lobe 1996; Schwechheimer, Smith, and Bevan 1998). Our data may explain this apparent discrepancy, as they suggest that polyglutamine repeats may be neutral in some proteins and not in others and that rapidly evolving repeats are more nearly neutral than conserved repeats. Searches for a functional role for polyglutamine repeats in proteins should therefore focus on proteins, such as those in our group B, that show conservation of Gln repeats over long periods of evolutionary time.

In conclusion, we suggest that the following interplay of forces influences the emergence of polyglutamine repeats. Glutamine repeats emerge preferentially in a sequence environment biased toward an overrepresentation of Gln codons (and possibly also related codons such as CCG). These concentrations occur in a class of proteins enriched in these codons by selection for a high content of Gln (as well as Pro, His, and Ser). Repeats emerge in regions of proteins that are subject to lower-than-average levels of purifying selection (Nishizawa, Nishizawa, and Kim 1999), as indicated by their nonsynonymous divergence rate, although the whole proteins are not subject to atypically low levels of purifying selection. We therefore propose that emerging repeats evolve as essentially neutral structures. As such, we would expect them to be gained or lost in a manner that reflects the underlying dynamics of the mutational process, thought to be predominantly replication slippage. Recent evidence suggests that slippage shows a bias toward expansion for short repeats coupled with shortening of longer repeats (Ellegren 2000; Xu et al. 2000), which would give rise to net expansion of new repeats. However, changes in the strength of purifying selection acting on the region of the protein containing the repeat may result in the repeat ceasing to be a neutral structure and becoming fixed in length, as appears to have happened in the proteins in our group B, which contain a repeat in both species. Fixation of repeats, or the susceptibility of proteins to incorporation of them, may reflect the general functional class of the protein concerned, as certain classes of proteins in *Saccharomyces cerevisiae,* notably transcription factors and protein kinases, are significantly enriched in Gln repeats (Albà, Santibáñez-Koref, and Hancock 1999*b*). If purifying selection plays an important role in regulating the emergence of CAG repeats in proteins, the recent suggestion that nonsynonymous substitution rates may vary systematically around mammalian genomes (Williams and Hurst 2000), perhaps reflecting variation in recombination frequency along chromosomes, may have implications for the chromosomal distribution of repeat-containing proteins.

Acknowledgments

We thank the U.K. Medical Research Council for support. E.A.W. received an MRC postgraduate studentship.

LITERATURE CITED

ABBOTT, C., and D. CHAMBERS. 1994. Analysis of CAG trinucleotide repeats from mouse cDNA sequences. Ann. Hum. Genet. **58**:87–94.

ALBÀ, M. M., M. F. SANTIBÁÑEZ-KOREF, and J. M. HANCOCK. 1999*a*. Conservation of polyglutamine tract size between mice and humans depends on codon interruption. Mol. Biol. Evol. **16**:1641–1644.

———. 1999*b*. Amino acid reiterations in yeast are overrepresented in particular classes of proteins and show evidence of a slippage-like mutational process. J. Mol. Evol. **49**:789–797.

ALBANESE, V., S. HOLBERT, C. SAADA et al. (14 co-authors). 1998. CAG/CTG and CGG/GCC repeats in human brain reference cDNAs: outcome in searching for new dynamic mutations. Genomics **47**:414–418.

ALTSCHUL, S. F., W. GISH, W. MILLER, E. W. MYERS, and D. J. LIPMAN. 1990. Basic local alignment search tool. J. Mol. Biol. **215**:403–410.

AMOS, W. 1999. A comparative approach to the study of microsatellite evolution. Pp. 66–79 *in* D. B. GOLDSTEIN and C. SCHLÖTTERER, eds. Microsatellites: evolution and applications. Oxford University Press, Oxford, England.

AOKI, M., L. KORANYI, A. C. RIGGS et al. (11 co-authors). 1996. Identification of trinucleotide repeat-containing genes in human pancreatic islets. Diabetes **45**:157–164.

BROCK, G. J. R., N. H. ANDERSON, and D. G. MONCKTON. 1999. Cis-acting modifiers of expanded CAG/CTG triplet repeat expandability: associations with flanking GC content and proximity to CpG islands. Hum. Mol. Genet. **8**:1061–1067.

BROHEDE, J., and H. ELLEGREN. 1999. Microsatellite evolution: polarity of substitutions within repeats and neutrality of flanking sequences. Proc. R. Soc. Lond. B Biol. Sci. **266**: 825–833.

BULLE, F., N. CHIANNILKULCHAI, A. PAWLAK, J. WEISSENBACH, G. GYAPAY, and G. GUELLAEN. 1997. Identification and chromosomal localization of human genes containing CAG/CTG repeats expressed in testis and brain. Genome Res. **7**: 705–715.

CHAMBERS, D. M., and C. M. ABBOTT. 1996. Isolation and mapping of novel mouse brain cDNA clones containing trinucleotide repeats, and demonstration of novel alleles in recombinant inbred strains. Genome Res. **6**:715–723.

DJIAN, P., J. M. HANCOCK, and H. S. CHANA. 1996. Codon repeats in genes associated with human diseases: fewer repeats in the genes of nonhuman primates and nucleotide substitutions concentrated at the sites of reiteration. Proc. Natl. Acad. Sci. USA **93**:417–421.

ELLEGREN, H. 2000. Heterogeneous mutation processes in human microsatellite DNA sequences. Nat. Genet. **24**:400–402.

ELLEGREN, H., C. R. PRIMMER, and B. C. SHELDON. 1995. Microsatellite ‘evolution’: directionality or bias? Nat. Genet. **11**:360–362.

GENETICS COMPUTER GROUP. 1997. Wisconsin package. Version 9.1. GCG GENETICS COMPUTER GROUP. 1997. Wisconsin package. Version 9.1. GCG, Madison, Wis.

GRAUR, D. 1985. Amino acid composition and the evolutionary rates of protein-coding genes. J. Mol. Evol. **22**:53–62.

GREEN, H., and N. WANG. 1994. Codon reiteration and the evolution of proteins. Proc. Natl. Acad. Sci. USA **91**:4298–4302.

HANCOCK, J. M., P. J. SHAW, F. BONNETON, and G. A. DOVER. 1999. High sequence turnover in the regulatory regions of the developmental gene hunchback in insects. Mol. Biol. Evol. **16**:253–265.

HEIN, J. J. 1990. Unified approach to alignment and phylogenies. Methods Enzymol. **183**:626–645.

HIGGINS, D. G., and P. M. SHARP. 1989. Fast and sensitive multiple sequence alignments on a microcomputer. Comput. Appl. Biosci. **5**:151–153.

JIANG, J. X., R. H. DEPREZ, E. C. ZWARTHOFF, and P. H. RIEGMAN. 1995. Characterization of four novel CAG repeat-containing cDNAs. Genomics **30**:91–93.

JUKES, T. H., and C. R. CANTOR. 1969. Evolution of protein molecules. Pp 21–132 *in* H. N. MUNRO, ed. Mammalian protein metabolism. Academic Press, New York.

JURKA, J., and C. PETHIYAGODA. 1995. Simple repetitive DNA sequences from primates: compilation and analysis. J. Mol. Evol. **40**:120–126.

KARLIN, S., and C. BURGE. 1996. Trinucleotide repeats and long homopeptides in genes and proteins associated with nervous system disease and development. Proc. Natl. Acad. Sci. USA **93**:1560–1565.

KAZEMI-ESFARJANI, P., M. A. TRIFIRO, and L. PINSKY. 1995. Evidence for a repressive function of the long polyglutamine tract in the human androgen receptor: possible pathogenetic relevance for the (CAG)n-expanded neuronopathies. Hum. Mol. Genet. **4**:523–527.

KIM, S. J., B. H. SHON, J. H. KANG, K. S. HAHM, O. J. YOO, Y. S. PARK, and K. K. LEE. 1997. Cloning of novel trinucleotide-repeat (CAG) containing genes in mouse brain. Biochem. Biophys. Res. Commun. **240**:239–243.

KING, B. L., G. SIRUGO, J. H. NADEAU, T. J. HUDSON, K. K. KIDD, B. M. KACINSKI, and M. SCHALLING. 1998. Long CAG/CTG repeats in mice. Mamm. Genome **9**:392–393.

KRUGLYAK, S., R. T. DURRETT, M. D. SCHUG, and C. F. AQUADRO. 1998. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proc. Natl. Acad. Sci. USA **95**: 10774–10778.

KUMAR, S., T. TAMURA, and M. NEI. 1993. MEGA: molecular evolutionary genetics analysis. Version 1.01. Pennsylvania State University, University Park.

LANZ, R. B., S. WIELANDS, M. HUG, and S. RUSCONI. 1995. A transcriptional repressor obtained by alternative translation of a trinucleotide repeat. Nucleic Acids Res. **23**:138–145.

LI, S. H., M. G. MCINNIS, R. L. MARGOLIS, S. E. ANTONARAKIS, and C. A. ROSS. 1993. Novel triplet repeat containing genes in human brain: cloning, expression, and length polymorphisms. Genomics **16**:572–579.

MAKALOWSKI, W., and M. S. BOGUSKI. 1998. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. Proc. Natl. Acad. Sci. USA **95**:9407–9412.

MARGOLIS, R. L., M. R. ABRAHAM, S. B. GATCHELL, S. H. LI, A. S. KIDWAI, T. S. BRESCHEL, O. C. STINE, C. CALLAHAN, M. G. MCINNIS, and C. A. ROSS. 1997. cDNAs with long CAG trinucleotide repeats from human brain. Hum. Genet. **100**:114–122.

MATASSI, G., P. M. SHARP, and C. GAUTIER. 1999. Chromosomal location effects on gene sequence evolution in mammals. Curr. Biol. **9**:786–791.

MICHALAKIS, Y., and M. VEUILLE. 1996. Length variation of CAG/CAA trinucleotide repeats in natural populations of Drosophila melanogaster and its relation to the recombination rate. Genetics **143**:1713–1725.

MORIN, P. A., P. MAHBOUBI, S. WEDEL, and J. ROGERS. 1998. Rapid screening and comparison of human microsatellite markers in baboons: allele size is conserved, but allele number is not. Genomics **53**:12–20.

MOUCHIROUD, D., C. GAUTIER, and G. BERNARDI. 1995. Frequencies of synonymous substitutions in mammals are gene-specific and correlated with frequencies of nonsynonymous substitutions. J. Mol. Evol. **40**:107–113.

NAKACHI, Y., T. HAYAKAWA, H. OOTA, K. SUMIYAMA, L. WANG, and S. UEDA. 1997. Nucleotide compositional constraints on genomes generate alanine-, glycine-, and proline-rich structures in transcription factors. Mol. Biol. Evol. **14**: 1042–1049.

NAKAMURA, Y., T. GOJOBORI, and T. IKEMURA. 2000. Codon usage tabulated from international DNA sequence databases: status for the year 2000. Nucleic Acids Res. **25**:244–245.

NERI, C., V. ALBANESE, A. S. LEBRE et al. (23 co-authors). 1996. Survey of CAG/CTG repeats in human cDNAs representing new genes: candidates for inherited neurological disorders. Hum. Mol. Genet. **5**:1001–1009.

NISHIZAWA, M., and K. NISHIZAWA. 1998. Biased usages of arginines and lysines in proteins are correlated with local-scale fluctuations of the G + C content of DNA sequences. J. Mol. Evol. **47**:385–393.

NISHIZAWA, K., M. NISHIZAWA, and K. S. KIM. 1999. Tendency for local repetitiveness in amino acid usages in modern proteins. J. Mol. Biol. **294**:937–953.

OHTA, T., and Y. INA. 1995. Variation in synonymous substitution rates among mammalian genes and the correlation between synonymous and nonsynonymous divergences. J. Mol. Evol. **41**:717–720.

PAWLAK, A., N. CHIANNIKULCHAI, W. ANSORGE, F. BULLE, J. WEISSENBACH, G. GYAPAY, and G. GUELLAEN. 1998. Identification and mapping of 26 human testis mRNAs containing CAG/CTG repeats. Mamm. Genome **9**:745–748.

PEARSON, W. R., and D. J. LIPMAN. 1988. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA **85**:2444–2448.

PINTO, M., and C. G. LOBE. 1996. Products of the grg (Groucho-related gene) family can dimerize through the amino-terminal Q domain. J. Biol. Chem. **271**:33026–33031.

REDDY, P. H., E. STOCKBURGER, P. GILLEVET, and D. A. TAGLE. 1997. Mapping and characterization of novel (CAG)n repeat cDNAs from adult human brain derived by the oligo capture method. Genomics **46**:174–182.

RIGGINS, G. J., L. K. LOKEY, J. L. CHASTAIN, H. A. LEINER, S. L. SHERMAN, K. D. WILKINSON, and S. T. WARREN. 1992. Human genes containing polymorphic trinucleotide repeats. Nat. Genet. **2**:186–191.

RUBINSZTEIN, D. C. 1999. Trinucleotide expansion mutations cause diseases which do not conform to classical Mendelian expectations. Pp. 80–97 *in* D. B. GOLDSTEIN and C. SCHLÖTTERER, eds. Microsatellites: evolution and applications. Oxford University Press, Oxford, England.

RUBINSZTEIN, D. C., B. AMOS, and G. COOPER. 1999. Microsatellite and trinucleotide-repeat evolution: evidence for mutational bias and different rates of evolution in different lineages. Philos. Trans. R. Soc. Lond. B Biol. Sci. **354**: 1095–1099.

RUBINSZTEIN, D. C., W. AMOS, J. LEGGO, S. GOODBURN, S. JAIN, S. H. LI, R. L. MARGOLIS, C. A. ROSS, and M. A. FERGUSON-SMITH. 1995*a.* Microsatellite evolution—evidence for directionality and variation in rate between species. Nat. Genet. **10**:337–343.

RUBINSZTEIN, D. C., W. AMOS, J. LEGGO, S. GOODBURN, R. S. RAMESAR, J. OLD, R. BONTROP, R. MCMAHON, D. E. BARTON, and M. A. FERGUSON-SMITH. 1994. Mutational bias provides a model for the evolution of Huntington's disease and predicts a general increase in disease prevalence. Nat. Genet. **7**:525–530.

RUBINSZTEIN, D. C., J. LEGGO, G. A. COETZEE, R. A. IRVINE, M. BUCKLEY, and M. A. FERGUSON-SMITH. 1995*b.* Sequence variation and size ranges of CAG repeats in the Machado-Joseph disease, spinocerebellar ataxia type 1 and androgen receptor genes. Hum. Mol. Genet. **4**:1585–1590.

SCHMID, K. J., and D. TAUTZ. 1999. A comparison of homologous developmental genes from Drosophila and Tribolium reveals major differences in length and trinucleotide repeat content. J. Mol. Evol. **49**:558–566.

SCHWECHHEIMER, C., C. SMITH, and M. W. BEVAN. 1998. The activities of acidic and glutamine-rich transcriptional activation domains in plant cells: design of modular transcription factors for high-level expression. Plant Mol. Biol. **36**: 195–204.

STALLINGS, R. L. 1994. Distribution of trinucleotide microsatellites in different categories of mammalian genomic sequence: implications for human genetic diseases. Genomics **21**:116–121.

TAUTZ, D., and L. NIGRO. 1998. Microevolutionary divergence pattern of the segmentation gene hunchback in Drosophila. Mol. Biol. Evol. **15**:1403–1411.

THOMPSON, J. D., D. G. HIGGINS, and T. J. GIBSON. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673–4680.

TICHER, A., and D. GRAUR. 1989. Nucleic acid composition, codon usage, and the rate of synonymous substitution in protein-coding genes. J. Mol. Evol. **28**:286–298.

TREIER, M., C. PFEIFLE, and D. TAUTZ. 1989. Comparison of the gap segmentation gene hunchback between Drosophila melanogaster and Drosophila virilis reveals novel modes of evolutionary change. EMBO J. **8**:1517–1525.

WILLIAMS, E. J. B., and L. D. HURST. 2000. The proteins of linked genes evolve at similar rates. Nature **407**:900–903.

XU, X., M. PENG, Z. FANG, and X. XU. 2000. The direction of microsatellite mutations is dependent upon allele length. Nat. Genet. **24**:396–399.

ZUHLKE, C., R. KIEHL, A. JOHANNSMEYER, K. H. GRZESCHIK, and E. SCHWINGER. 1999. Isolation and characterization of novel CAG repeat containing genes expressed in human brain. DNA Seq. **10**:1–6.

DIETHARD TAUTZ, reviewing editor

Accepted January 29, 2001