J Mol Evol (2001) 52:249–259
DOI: 10.1007/s002390010153

JOURNAL OF MOLECULAR EVOLUTION

The Comparative Genomics of Polyglutamine Repeats: Extreme Difference in the Codon Organization of Repeat-Encoding Regions Between Mammals and *Drosophila*

M. Mar Albà,[1] Mauro F. Santibáñez-Koref,[2] John M. Hancock[2,*]

[1]Wohl Virion Centre, Windeyer Institute of Medical Sciences, University College London, London W1P 6DB, U.K.
[2]Comparative Sequence Analysis Group, MRC Clinical Sciences Centre, Imperial College School of Medicine, Hammersmith Hospital, London W12 0NN, U.K.

Received: 17 August 2000 / Accepted: 20 November 2000

Abstract. Polyglutamine repeats within proteins are common in eukaryotes and are associated with neurological diseases in humans. Many are encoded by tandem repeats of the codon CAG that are likely to mutate primarily by replication slippage. However, a recent study in the yeast *Saccharomyces cerevisiae* has indicated that many others are encoded by mixtures of CAG and CAA which are less likely to undergo slippage. Here we attempt to estimate the proportions of polyglutamine repeats encoded by slippage-prone structures in species currently the subject of genome sequencing projects. We find a general excess over random expectation of polyglutamine repeats encoded by tandem repeats of codons. We nevertheless find many repeats encoded by nontandem codon structures. Mammals and *Drosophila* display extreme opposite patterns. *Drosophila* contains many proteins with polyglutamine tracts but these are generally encoded by interrupted structures. These structures may have been selected to be resistant to slippage. In contrast, mammals (humans and mice) have a high proportion of proteins in which repeats are encoded by tandem codon structures. In humans, these include most of the triplet expansion disease genes.

Key words: Glutamine repeats — Replication slippage — Comparative genome analysis — Repeat evolution — Triplet expansion diseases — Triplet repeats — Genome evolution

Introduction

Amino acid repeats, particularly of uncharged polar or acidic amino acids, are much more common in eukaryotic proteins that can be explained by the frequencies of the individual amino acids (Green and Wang 1994; Karlin and Burge 1996). These structures are of evolutionary interest because there is some evidence that they can mediate or modulate protein–protein interactions (Mitchell and Tjian 1989; Perutz et al. 1994; Kazemi-Esfarjani et al. 1995), raising the possibility that changes in their length during evolution could result in changes in the strength of protein–protein interactions. As a number of studies (Wharton et al. 1985; Gerber et al. 1994; Bhandari and Brahmachari 1995; Karlin and Burge 1996; Nakachi et al. 1997; Hancock 1993; Wilkins and Lis 1999; Agianian et al. 1999; Xiao and Jeang 1998; Alba et al. 1999a), including a whole-genome analysis in *Saccharomyces cerevisiae* (Alba et al. 1999a), have associated glutamine repeats with transcription factors, this could have implications for the evolution of gene regulatory networks (Hancock 1993; Karlin and Burge 1996; Richard and Dujon 1997). Interest in the functional con-

*Present address: Department of Computer Science, Royal Holloway University of London, Egham, Surrey TW20 0EX, U.K.
Correspondence to: Dr. John M. Hancock; e-mail: J.Hancock@dcs.rhul.ac.uk

text of such repeats has also been stimulated by the association of several proteins that contain abnormally expanded glutamine tracts with human neurological disorders (reviewed by Reddy and Housman 1997).

To understand the evolutionary origins of polyglutamine repeats it is important to understand the mutational processes that are most important in shaping them. In principle, regions encoding glutamine repeats can arise by two types of process—the sequential expansion of a short codon repeat by a process such as replication slippage (Levinson and Gutman, 1987) and the accumulation of point mutations in the relevant part of a gene (Alba et al. 1999a). There is no evidence for a role for recombination in the evolution of this class of tandem repeats (reviewed by Sia et al. 1997) but some slippage-like processes, for example, the formation of hairpin loops on the replicating strands (e.g., Ohshima and Wells 1997), could contribute to their mutation and evolutionary expansion. Evolutionary studies of a number of genes involved in the human triplet expansion diseases have indicated that the repeats in these genes have arisen by a gradual expansion of the tandem repeat (Rubinsztein et al. 1994, 1995; Djian et al. 1996; Pecheux et al. 1996; Limprasert et al. 1996, 1997; Choong et al. 1998; R. Gangeswaran, HS Chana, M.F.S.-K. and J.M.H., manuscript in preparation) apparently resulting from replication slippage. However, this cannot be generalized to all such repeats, as these genes may be atypical because of their high mutation rate and association with disease.

Rather than considering individual genes, an alternative way to study this question is to take a whole-genome approach or as near to such an approach as is practically possible (Alba et al. 1999a,b). Using database screens it can be shown that glutamine repeats in human and mouse proteins fall into two distinct classes, which are encoded either by pure or nearly pure repeats of a single codon (usually CAG) or by more or less random mixtures of codons (Alba et al. 1999b). Structures with an intermediate level of organization (i.e., containing a number of pure runs interspersed by interrupting codons) are relatively rarer. Tandemly repetitive sequences are prone to replication slippage (Levinson and Gutman 1987), while, although slippage-like patterns of change can be seen even in cryptically repetitive sequence regions (Hancock and Vogler 2000) and slippage can be modeled even for sequences that are not tandemly repetitious (Jones and Kafatos 1982), rates of slippage mutation are substantially lower at tandem repeats that are interrupted by point mutations (e.g., Rolfsmeier and Lahue 2000). These two extreme classes may therefore represent a class that has arisen primarily by replication slippage, which gave rise to pure codon repeats, and a class in which replication slippage is relatively less prominent. Thus different glutamine repeats, in different genes, may arise essentially by the different mechanisms outlined above. An analysis of the complete set of sequences encoding polyglutamine repeats in the yeast genome (Alba et al. 1999a) indicated that the majority does not consist of long runs of single codons, suggesting that in yeast point mutation is an important process in generating polyglutamine repeats.

These observations raise the question to what extent the contribution of point mutation and slippage to the evolution of these structures differs in different evolutionary lineages. To study this we have analyzed large protein data sets from a further four model organisms that are currently the subjects of genome sequencing projects (*Escherichia coli, Caenorhabditis elegans, Arabidopsis thaliana, Drosophila melanogaster*) and compared them with *S. cerevisiae, Mus musculus,* and *Homo sapiens* repeats. The results show similarities and differences between species. For most of the eukaryotic species there is an overrepresentation of tracts encoded by long CAG tandem repeats, supporting the idea that recent slippage has been involved in the generation of a significant proportion of the tracts. However, on average about 70% of the tracts do not show evidence of recent slippage, and in *D. melanogaster* there is no clear evidence of a strong contribution from slippage. Furthermore, in the two mammalian species about one-third of the tracts are exclusively encoded by CAG and the length of the tracts is on average much longer than in other species. This suggests that slippage has played a more important role in the evolution of polyglutamine regions in mammals than in other taxa.

Methods

Database Searches

BLASTP (Altschul et al. 1990) at the NCBI was used to find all GenBank entries which contained genes encoding long polyglutamine tracts (≥6 glutamines) from *E. coli, S. cerevisiae, C. elegans, A. thaliana, D. melanogaster, M. musculus,* and *H. sapiens.* Redundancy in the primary data sets was eliminated by running FASTA within the GCG package (Pearson and Lipman 1988; GCG 1997). Sequences with 95% identity were considered redundant, and only one representative sequence was used in the subsequent analysis. Where there was a discrepancy in the length of the polyglutamine tract in nearly identical sequences, we took the sequence with the longest tract.

Analysis of Codon Repeats

We used statistical analysis to analyze two properties of polyglutamine repeat-encoding regions. The first was the extent of deviation of the codon organization within these regions from random. This was measured by considering the deviation of the length of the longest run of each codon type from chance expectation (Alba et al. 1999a,b). The second property was the over- or underrepresentation of tandem codon repeats of a particular length in the whole set of polyglutamine-coding regions in a given species.

Length of the Longest Homogeneous Run. As described previously (Alba et al. 1999a,b) the organizational homogeneity or otherwise of a region encoding a polyglutamine repeat has to be considered in the

Table 1. Polyglutamine tracts in different species

			Length of polyglutamine tract				CAG relative frequency[a]		Pure codon tracts	
Species	Total	Proteins	6–7	8–10	>10	Max.	Genome	Tracts	CAG	CAA
S. cerevisiae	147	117	37%	28%	35%	37	0.307	0.450*	4.7%	5.4%
C. elegans	137	118	70%	26%	4%	14	0.331	0.430*	2.2%	5.8%
A. thaliana	44	37	70%	16%	14%	17	0.442	0.465 (NS)	2.2%	11.3%
D. melanogaster	312	288	52%	21%	22%	28	0.716	0.728 (NS)	7.3%	0%
M. musculus	56	28	37%	23%	40%	37	0.743	0.824*	37.2%	0%
H. sapiens	96	68	36%	19%	42%	40	0.674	0.830*	26.2%	0%

[a] Chi-square test of the relative number of CAG codons in each tract compared to the expectation from the CAG genome frequency: * $p < 0.0001$; NS, not significant.

context of a random expectation, based on the probability of finding the two types of glutamine codons in the region, and the region's length. We have measured this by calculating Z statistics (L_p and L_m) that measure the extent to which the organization of codons in such a region deviates from chance expectation. These measures describe the length of the longest homogeneous run of a particular codon in an array. L_p, which is formally $Z(l|g,p)$ (Alba et al. 1999a, b), is the standardized deviation of the length of the longest run of a particular codon in a given array from the expected length, based on the overall genome codon usage. The equivalent measure making use of the codon frequencies in the array itself, L_m [formally $Z(l|g,m)$], was defined as

$$L_m = \frac{(l - E(l|g,m))}{\sqrt{V(l|g,m)}}$$

Here g is the length of the array, m the number of CAG or CAA codons in the array, $E(l|g,m) = \Sigma_{j=0}^{g}\, jp(j|g,m)$ the expected length of CAG or CAA runs, and $V(l|g,m) = \Sigma_{j=0}^{g}\, j^2p(j|g,m) - (E_i(l|g,m))^2$ the variance. The variance and the expected value can be determined from $p_l(l|g,m)$, the probability that the longest run of a particular codon is l, given that the codon occurs m times in an array of length g. This can be calculated as

$$p_l\,(l|g,m) = \frac{1}{\binom{g}{m}} \sum_{r=1}^{m} N_l(m,r,l)\, \frac{N_k(g,m,r)}{\binom{m-1}{r-1}}$$

where

$$N_l(m,r,l) = \sum_{i=1}^{r} (-1)^i \binom{r}{i} \left(\binom{m - li - 1}{r-1} - \binom{m - (l-1)i - 1}{r-1} \right)$$

and

$$N_k(g,m,r) = \binom{m-1}{r-1} \left(\binom{g-m-1}{r} + 2\binom{g-m-1}{r-1} + \binom{g-m-1}{r-2} \right)$$

Number of Runs of a Given Length. To assess whether homogenous runs of length l occur more often than expected, we defined a second measure, $S(1)$ [formally $Z_n(n|g,p,l)$]. The corresponding probability $p_s(l|g,p,l)$, the probability of finding n homogeneous runs of length l in an array, given a genomic codon usage p for a particular codon, was calculated as

$$p_s\,(n|g,p,l) = \frac{1}{\binom{g}{p}} \sum_{r=1}^{p} N_s(p,r,n,l)\, \frac{N_k(g,p,r)}{\binom{p-1}{r-1}}$$

where

$$N_s(p,r,n,l) = \sum_{i=n}^{r} (-1)^{i-n} \binom{i}{n} \binom{p - li - 1}{r - i - 1} \binom{r}{i}$$

and $N_k(g,p,r)$ is defined as before.

Genomic codon frequencies were obtained from the species Codon Usage Table from the CUTG database at http://www.kazusa.or.jp/codon/ (Nakamura et al. 2000).

Results

General Characteristics of Polyglutamine Repeats in Different Species

A search for glutamine repeats of length 6 or greater was performed for seven species: *E. coli*, *S. cerevisiae*, *C. elegans*, *D. melanogaster*, *A. thaliana*, *M. musculus*, and *H. sapiens.* No proteins from *E. coli* contained glutamine homopeptides of this length but between 28 (for mouse) and 312 (for *D. melanogaster*) were found for the eukaryote species (Table 1).

Three main properties of these repeats were first analyzed.

Length. There are striking differences in the length of polyglutamine regions in different species. The longest glutamine tracts are found in the mammalian species, with 40% of the mouse tracts and 45% of the human ones being longer than 10 glutamines. On the other hand, the tracts in *C. elegans* were noticeably shorter, only 4% being longer than 10 glutamines (Table 1). The longest repeat was found in a human protein, the uncharacterized cDNA CAGH44 (Margolis et al., 1997), which contains a tract of 40 glutamines (Table 1). Tracts of nearly the same length, 37 glutamines, were detected in *M. musculus* and *S. cerevisiae.*

Codon Content. The relative contributions of CAG and CAA codons were calculated for the repeats of each species. When these codon frequencies were compared to the species codon usage (Nakamura et al. 2000), we observed an overrepresentation of CAG codons in *S. cerevisiae, C. elegans, M. musculus,* and *H. sapiens* repeats but no significant deviations from expectation for *D. melanogaster* and *A. thaliana* (Table 1).

Pure Trinucleotide Repeats. Homopeptides encoded exclusively by CAG were relatively rare in the nonmammalian species but were common in mammals, making up 37% of mouse polyglutamine tracts and 26% of human ones (Table 1). These differences cannot be explained solely by differences in codon frequency. For example, *D. melanogaster* has an overall CAG frequency similar to that of *M. musculus* and *H. sapiens* but contained a much lower percentage of tracts made up exclusively by this codon (Table 1).

Analysis of the Codon Arrangement in Regions Encoding Polyglutamine Tracts

If expansion of trinucleotides by slippage had given rise to a polyglutamine tract recently in evolution, we should observe an excess of perfect repeats of either CAA or CAG in the region encoding the homopeptide. On the contrary, if slippage had not been the main process giving rise to a polyglutamine tract, or had not occurred recently, we should observe an arrangement of the two codons more in accordance with a random distribution. We have characterized the relative balance of the two processes by measuring two values, which we have called L_p and L_m. Both of these values measure, for individual regions coding for glutamine repeats, whether they contain a longer run of codons than would be expected by chance. The difference between the measures is that L_p calculates the expected length of the longest codon run in an array based on the relative frequencies of CAG and CAA in the total set of protein coding sequences in a particular genome. L_m, on the other hand, uses the frequencies in the region under study. L_m has the advantage that it does not assume that the relative uses of the two codons should correspond to the genomic average but has the disadvantage that it cannot be used for regions made up of a single codon, as in these cases the method will inevitably predict only a single homogeneous array of the length observed. In the case where slippage has made a major contribution to the expansion of codon repeats in regions coding for glutamine repeats, we would expect to see an excess (>5%) of arrays with L_p and L_m values greater than 1.645 (called the 95% interval here). If slippage had been essentially absent, we would expect no excess. The regions encoding polyglutamine tracts from different species showed important differences in the frequency distribution of L_m and L_p values, summarized in Table 2.

Table 2. Deviations from randomness in the lengths of runs of single codons

	CAG[a]				CAA				
	<5%		>95%		<5%		>95%		
Species	L_p	L_m	L_p	L_m	L_p	L_m	L_p	L_m	Total[b]
S. cerevisiae	3	*2*	31	*14*	12	*6*	8	*10*	147
C. elegans	3	*0*	26	*9*	2	*3*	7	*8*	137
A. thaliana	11	*3*	27	*14*	2	*0*	23	*8*	44
D. melanogaster	2	*3*	8	*8*	0	*0*	7	*8*	312
M. musculus	0	*7*	33	*7*	18	*0*	4	*7*	56
H. sapiens	1	*0*	25	*10*	15	*0*	3	*3*	96

[a] Number of repeats scoring less than −1.64 (<5%) or more than 1.64 (>95%) for the two measures L_p and L_m.
[b] Total number of repeats considered for each species.

In all eukaryotic species more than 5% of CAG arrays lay in the 95% interval, and in five of the six eukaryotes this proportion was between 25 and 34% (Table 2). This indicates a considerable excess of expanded CAG runs in these species. In contrast, a lower percentage of *D. melanogaster* tracts falls in this interval (8%). Only in two cases (L_p for *A. thaliana* and L_m for *M. musculus*) was there an excess of values below −1.645. In general, the proportions in the 95% interval were greater for L_p than for L_m, reflecting the high abundance of tracts encoded by a single codon, especially in mouse and human (Table 1).

When L_p and L_m were calculated for CAA, the percentage of arrays in the 95% interval was much smaller, especially comparing the L_p proportions with those obtained previously for CAG. Such low values indicate that there is not a substantial excess of long pure CAA runs. The exception to this was *A. thaliana,* in which 22% of tracts lay in the 95% interval of the L_p distribution for CAA, compared to 27% for CAG. Both CAG and CAA perfect tandem repeats therefore tend to be abnormally long in *A. thaliana.*

Special Case No. 1: No Evidence of Long CAG Reiterations Encoding Polyglutamine Tracts in Drosophila melanogaster

A number of features of the regions encoding *D. melanogaster* polyglutamine tracts indicate that they differ substantially from those of the other eukaryotic species. The *D. melanogaster* genome contains a high relative frequency of CAG, similar to that found in humans and mouse (Table 1). In contrast to these species, however, the percentage of pure CAG tracts is low (7.3 vs. 26.2 and 37.2%, respectively), even though polyglutamine tracts in *D. melanogaster* tend to be shorter than in the mammalian species and so are statistically more likely to be encoded by only a single codon type. Moreover, there is not a significant deviation in the number of CAG

codons in the reiterants from the overall *Drosophila* codon usage, in contrast to the observations in most of the other species (Table 1).

Another difference is found in the distribution of L_p for CAG. As described above, the proportion of arrays with L_m and L_p values in the 95% interval was low (8%) and close to the expectation under a random codon distribution (5%), contrary to the other eukaryotic species studied. The histogram of L_p for CAG reveals some interesting features (Fig. 1). Apart from a slight excess of tracts in the rightmost part of the graph, the main characteristic is that the values have a mean close to zero (Table 3) and show an excess of values close to the mean compared to the fitted normal distribution. This pattern is not observed in the other species, which show much higher means (Table 3) and a frequency peak close to or below the maximum of the fitted peak. The population of repeats in *D. melanogaster* is therefore highly homogeneous, the simplest explanation being that the regions encoding most of the tracts contain only short runs of CAG. To test for over- or underrepresentation of CAG runs of particular sizes, we developed another statistic, $S(l)$, where l is the size of the pure run, which is calculated for each run length. $S(l)$ measures the over- or underrepresentation of repeats of a given length in the data set compared to expectations based on genomic codon frequencies. Analysis of *D. melanogaster* tracts using this statistic shows an excess of CAG repeats of size 3 codons, as indicated by a significantly high average $S(3)$ value ($p < 0.001$). Tracts of sizes 1 and 2 are underrepresented ($p < 0.01$), and from size 3 onward they gradually become less abundant, with another minimum at size 10 (Fig. 1B). Similar features were not seen for other species.

Another feature of *D. melanogaster* tracts is the relationship between the values of L_p for CAG and the length of the polyglutamine tract. In the other species, values of L_p are similar in tracts of different length and are independent of the length of the homopeptide (not shown). In *D. melanogaster*, however, there is a significant negative correlation of L_p for CAG with the homopeptide length ($r = -0.21$, $p < 10^{-3}$) (Fig. 1C). This indicates that as the size of the polyglutamine tract increases, there is an effective decrease in the size of the longest perfect CAG repeat. Thus, glutamine repeats show little evidence of having been generated by slippage in *Drosophila.*

Special Case No. 2: Mammalian Polyglutamine Tracts Are Long and CAG Rich

About one-third of the polyglutamine tracts of *M. musculus* and *H. sapiens* in the GenBank database are formed exclusively by CAG repeats. This percentage is much higher than in other species, in which such tracts constitute only between 2.2 and 7.3% of the total (Table 1). The abundance of pure CAG repeats is reflected in the distribution of L_p for CAG (Table 2). A total of 24 human (25%) and 33 mouse (59%) arrays falls into the 95% interval of L_p. For L_m this deviation was much lower (Table 2), indicating that the major contribution to these high L_p sequences is from perfect CAG repeats that extend throughout the glutamine tracts. The histograms of L_p for human and mouse have a bimodal shape (Figs. 2A and B, respectively) not observed in the other species. Interestingly, in these histograms one of the two peaks is approximately centered at 0, as expected for a random distribution of CAG and CAA, while the other represents pure or almost-pure arrays (positive values). So, while a proportion of the tracts does not show deviations with respect to a random distribution of the two codons, another group contains significantly long CAG repeats.

Of 10 tracts found within human disease-associated proteins, 8 fall into the 95% interval of L_p for CAG, representing 33% of the total of tracts in this interval (Table 4). The strong interest in these genes has probably resulted in an overrepresentation of sequences with long CAG repeats in both the human and the mouse sequence databases, as extensive screening exercises for cDNAs containing long CAG repeats have been carried out in both species. This could lead to an overestimation of the extent to which codon reiteration occurs in human coding regions. The size of any such ascertainment bias was investigated in two ways. First, we asked if there was a significant correlation between L_p or L_m and array size, as ascertainment bias of this sort would be expected to enrich the sample in long, uniform CAG arrays which would have high L_p values. No significant correlation was found. To investigate whether the class of sequences with high L_p scores seen in mammals was characteristic of long repeats, we also compiled frequency histograms for short ($n < 10$) and long ($n > 10$) repeats. These histograms show more pronounced second maxima for short than for long repeats (Fig. 3). The maximum value of L_p is constrained by the number of possible arrangements of codons and therefore by the array length. Because of this, this second maximum could reflect an artifactual constraint on the value of L_p. To overcome this, we carried out a chi-square analysis of the number of repeats falling above and below 1.0 (which is effectively the border between the two classes in the histograms) for long and short repeats using a 2×2 contingency table. There was no significant deviation from randomness in this table in either mouse ($p = 0.54$) or human ($p = 0.27$). We therefore conclude that the secondary peaks we observe for L_p in the mammals are not due primarily to ascertainment biases. The true extent to which pure CAG repeats contribute to the generation of polyglutamine segments in mammalian proteins will nevertheless become clear only when not only the complete genome sequences but also high-quality annotation of coding regions is available for these species.

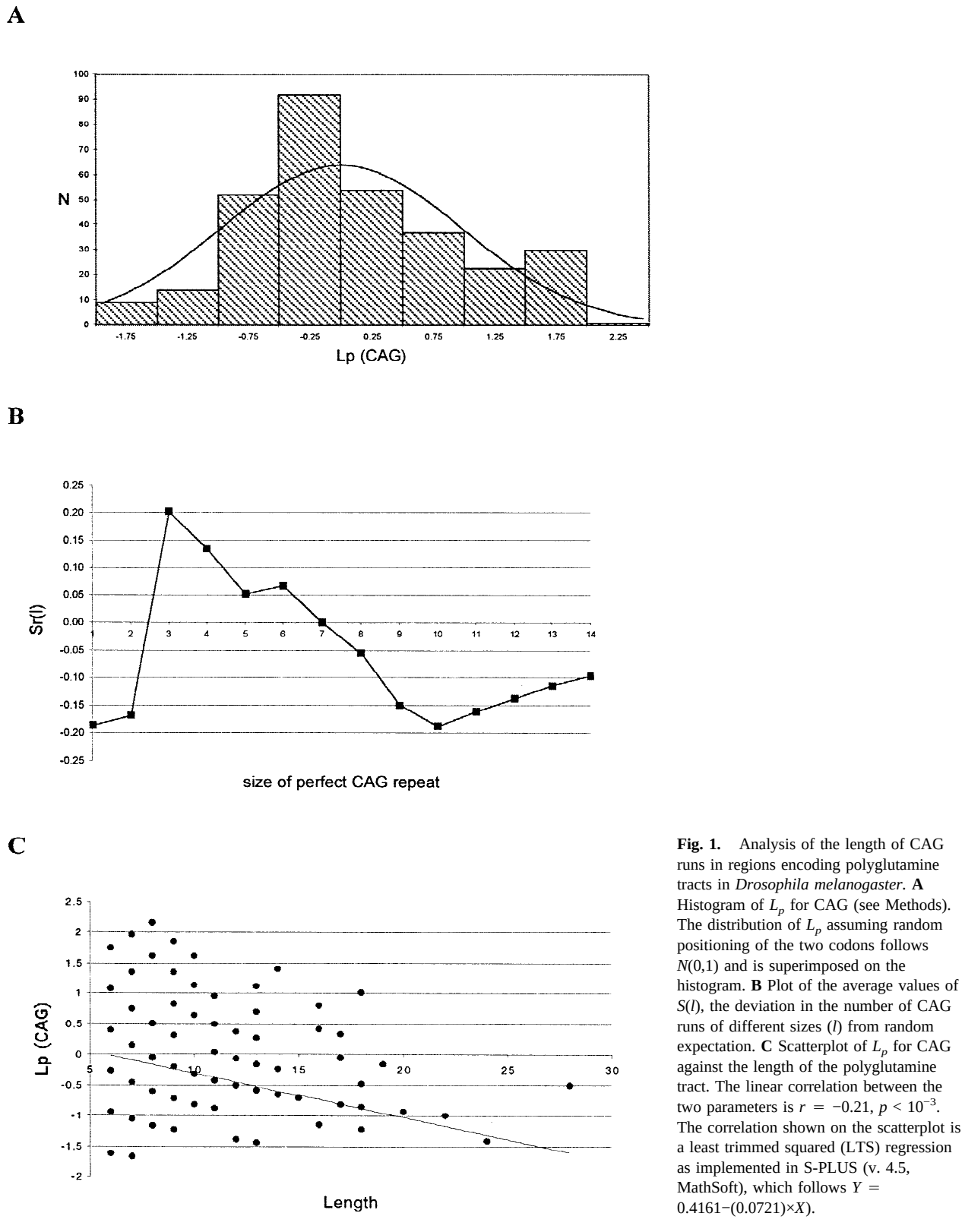

Fig. 1. Analysis of the length of CAG runs in regions encoding polyglutamine tracts in *Drosophila melanogaster*. **A** Histogram of L_p for CAG (see Methods). The distribution of L_p assuming random positioning of the two codons follows $N(0,1)$ and is superimposed on the histogram. **B** Plot of the average values of $S(l)$, the deviation in the number of CAG runs of different sizes (l) from random expectation. **C** Scatterplot of L_p for CAG against the length of the polyglutamine tract. The linear correlation between the two parameters is $r = -0.21$, $p < 10^{-3}$. The correlation shown on the scatterplot is a least trimmed squared (LTS) regression as implemented in S-PLUS (v. 4.5, MathSoft), which follows $Y = 0.4161-(0.0721)\times X$).

Discussion

These analyses show that while slippage has been important in generating many polyglutamine regions in eukaryotic proteins, differences are found between species in the proportions of glutamine repeats that show evidence of having been generated by slippage in the sense that they contain longer pure codon repeats than expected by chance. The species with the most prominent set of glutamine repeats showing evidence of slippage

Table 3. Estimated mean values of L_p and L_m for each species

Species	$L_p \pm$ SE[a]	$L_m \pm$ SE
C. elegans	0.780 ± 0.140	0.268 ± 0.080
D. melanogaster	0.091 ± 0.049	0.083 ± 0.057
A. thaliana CAG	0.555 ± 0.252	0.477 ± 0.154
A. thaliana CAA	0.367 ± 0.234	0.423 ± 0.151
S. cerevisiae	1.267 ± 0.193	0.144 ± 0.097
M. musculus	0.931 ± 0.227	0.065 ± 0.117
H. sapiens	0.854 ± 0.136	0.157 ± 0.099

[a] Values are given for CAG only except for *A. thaliana,* for which the CAA value is also given.

are the two mammals, human and mouse. These species also show two prominent and distinct classes of glutamine repeat. In contrast, *D. melanogaster* represents an opposite extreme in which there is no evidence of significant reiteration of single codons. Thus there appear to have been different relative levels of activity of slippage and point mutation in these regions in different evolutionary lineages.

Two kinds of influence could give rise to these extreme differences—differences in mutational processes and different selective conditions. Relative rates of slippage and point mutation might be altered by changes in the efficiency of mismatch repair and proofreading during DNA replication. This might affect slippage rates in particular, as there are known to be differences between species in the abilities of their mismatch repair machinery to detect loops of different length resulting from slippage during replication (reviewed by Eisen 1999). Another potential factor is differences in genome structure or organization between species. Long human CAG repeats introduced into transgenic mice tend to show much less instability than they do in humans (see Seznec et al. 2000). This appears to be a function of the sequence context in which sequences find themselves, as transgenic mice with larger segments of human flanking sequence behave in a more "human-like" manner (Seznec et al. 2000). This may reflect differences in simple properties such as base composition (Brock et al. 1999) or more complex properties such as local chromatin structure. It is perhaps noteworthy that the presence of two major classes of glutamine repeat (in mammals) coincides with the presence of isochores in the host genome (Bernardi 2000) in this limited data set. We cannot therefore exclude the (testable) possibility that the relative contributions of slippage-like and point mutational properties to polyglutamine repeat evolution reflect at least in part their local isochore environment. Differences in activity of slippage may also contribute to the origin and maintenance of isochores in these species.

Other mechanistic biases also probably come into play. The occurrence of longer than expected tracts of CAG is much more common than for CAA. This may indicate that the former has a higher tendency to undergo slippage. Consistent with this, tandem repeats of CAG are the most frequent trinucleotide microsatellites in human coding regions (Stallings 1994). An exception to the predominance of CAG over CAA repeats is found in *A. thaliana,* in which both CAG and CAA codons show a similar tendency to reiterate. Biases in the propensity of different motifs to undergo slippage have previously been suggested to be an important factor in determining the different frequencies of repetitive sequences found in a variety of genomic sequences (Hancock 1995). There is also some evidence that the processes giving rise to codon reiteration differs between species. In *Drosophila,* we observe an excess of three-codon repetition and relatively little one- or two-codon repetition. In *Arabidopsis* and mouse we find evidence of a significant excess of tracts which are highly interrupted. These observations may reflect differences in the capabilities of the repair mechanisms in these species or involvement of recombination (unequal crossing-over) as well as or instead of slippage.

Changes in the relative frequencies of slippage and point mutation might have a direct effect on the accumulation of long CAG repeats in genomes, as the length distribution of microsatellites in a genome has been suggested to reflect this balance, a higher relative rate of slippage giving rise to longer microsatellites (Kruglyak et al. 1998). However, polyglutamine repeats lie within protein coding regions and are therefore likely to be affected by selection as well as mutational processes. Genes containing glutamine repeats that are more conserved in length between human and mouse tend to show low nonsynonymous substitution rates, while genes containing more evolutionarily labile repeats tend to have higher rates (J.M.H., EA Worthey, and M.F.S.-K., submitted). This indicates that the level of selection acting on a gene containing a repeat has a significant impact on the repeat's evolution. Effects of selection could include disfavoring uninterrupted structures, as these are more likely to change in length with phenotypic consequences, such as triplet expansion disease in humans.

Homogeneous CAG repeats evolve in length more rapidly than interrupted repeats (Alba et al. 1999b). The bimodality of the distribution of L_p values for CAG in mammals suggests separate classes of proteins undergoing rapid and slow evolution of these repeat regions. As the rate of evolutionary length change of CAG repeats correlates positively with the rate of accumulation of nonsynonymous sequence changes (J.M.H., EA Worthey, and M.F.S.-K., submitted), mammalian glutamine repeat-containing proteins may divide into a rapidly evolving group, in which the repeats are evolutionarily neutral, or nearly so, and a conserved group in which the repeat length is constrained by evolution, perhaps giving rise to the high degree of interruption of codon organization observed in these proteins (Alba et al. 1999b).

Interestingly, *Drosophila* proteins that contain poly-

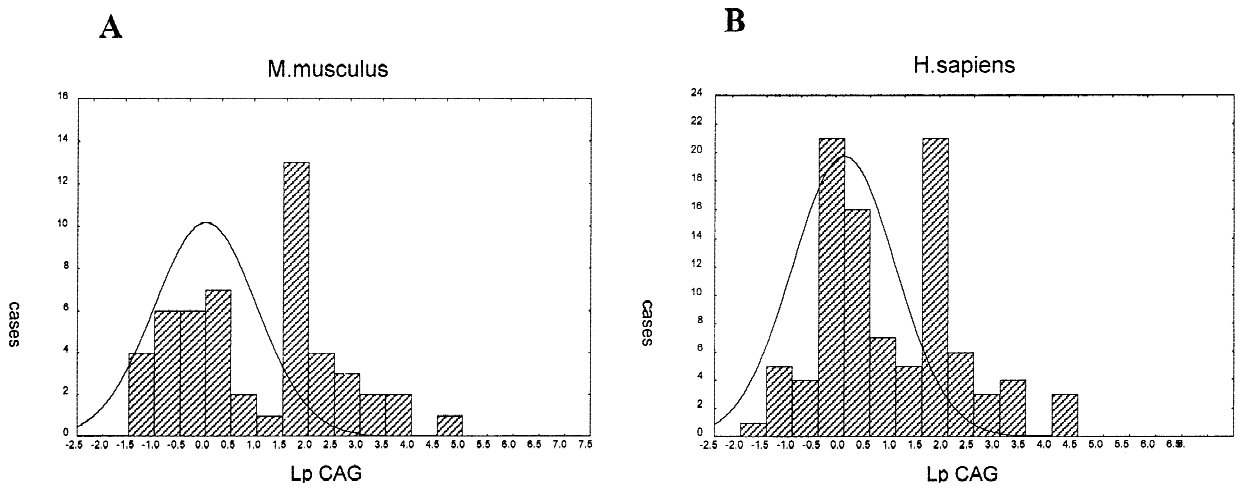

Fig. 2. Histogram of L_p for CAG for human **(A)** and mouse **(B)** polyglutamine tracts. Superimposed is the distribution assuming random positioning of the two codons, which follows the normal distribution [$N(0,1)$].

Table 4. Human glutamine homopeptides with $L_p > 1.645$ for CAG

	Length	Codon arrangement	Accession No.	Description of protein
Neurodegenerative disease genes	10	G10	AF020276	Spinocerebellar ataxia 7 (SCA7)
	12	G12	U79667	α1A-voltage-dependent calcium channel (SCA6)
	12	G12	X79204	Spinocerebellar ataxia 1 (SCA 1)
	14	G1A1G1A1G10	D38529	DRPLA
	15	G15	X79204	Spinocerebellar ataxia 1 (SCA1)
	21	G20A1	M20132	Androgen receptor (AR)
	23	G1A1G21	U64822	Josephin Machado-Joseph disease (MDJ1)
	23	G21A1G1	L12392	Huntington (Huntington disease; HD)
Nervous system	7	G7	U80735	CAGF28
	8	G8	U80745	CTG7a
	16	A2G1A1G12	U80745	CTG7a
	17	G2A2G13	U80750	CTG26
	18	A1G17	AB007945	KIAA0476
	30	G7A1G15A1G4A1G1	U80743	CAGH32
Transcription	10	G10	U49020	Myocyte-specific enhancer factor 2A (MEF2A)
	11	G11	U91935	Retina-derived POU-domain factor-1
	12	G12	L40992	Core-binding protein α subunit 1
	14	G14	L08424	Achaete-scute homologue
	15	G12A1G2	D84103	Mitochondrial DNA polymerase γ
	23	G1A2G3A1G13A1G2	D26155	Transcriptional activator hSNF2a
	38	G3A3G9A1G1A1G18A1G1	M55654	TATA-binding protein
Others	7	G7	Y08267	AAD10
	7	G7	Z72499	(HAUSP) herpesvirus-associated ubiquitin-specific protease
	7	G7	U07802	Tis11d, early response gene family
	14	G14	AF031815	Calcium-activated potassium channel (KCNN3)

glutamine tracts seem to be particularly common (N = 312). This suggests a different mode of evolution of polyglutamine-coding regions in this species. The observed patterns might be explained in a number of ways. First, slippage may be less active in *Drosophila* [as indicated by the low mutation rate of microsatellites (Schug et al. 1997)]. *Drosophila* shows a maximal value of $S(l)$ at repeat length three. $S(l)$ measures the deviation from expectation of the occurrence of CAG (or CAA) repeats of a particular length in the data set. Thus it is possible that slippage of 3-base pair motifs is controlled efficiently during DNA replication in *Drosophila* but

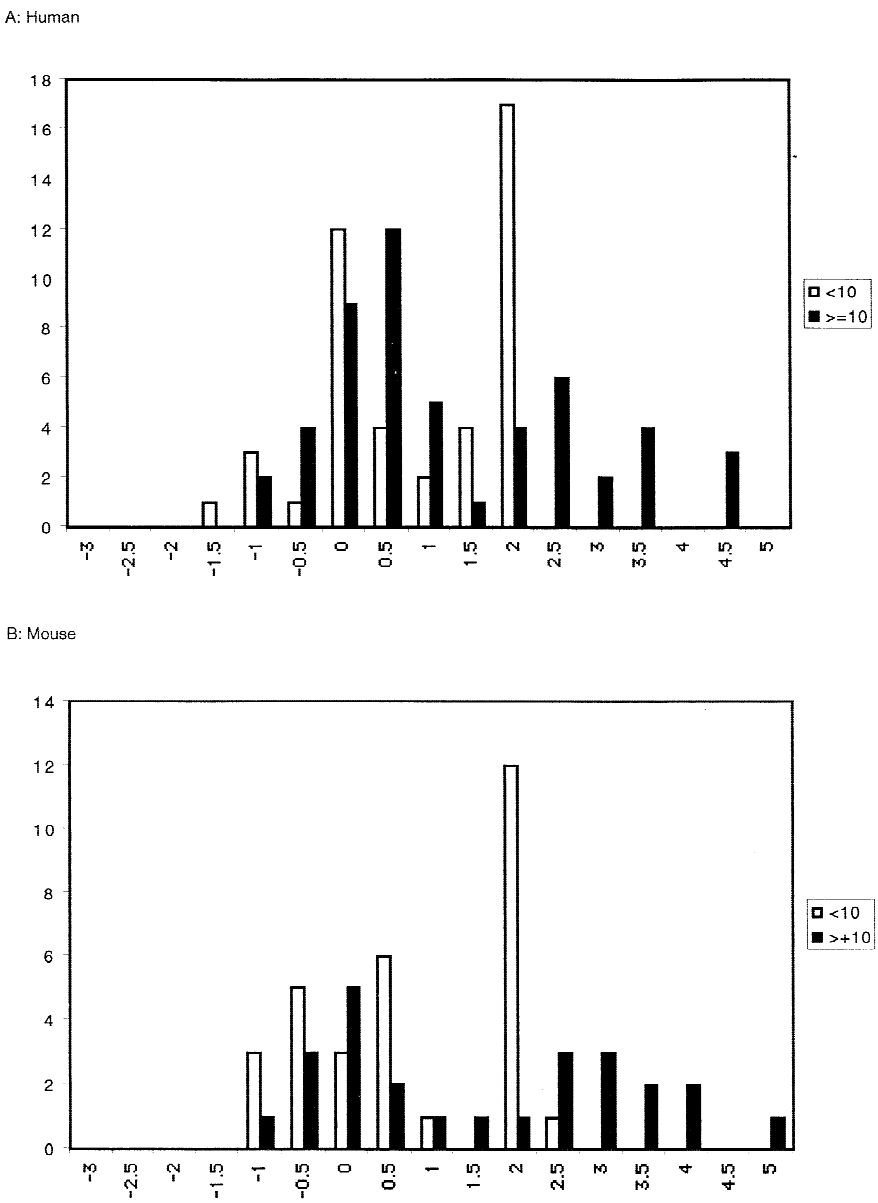

Fig. 3. Histograms of L_p for human and mouse, with data divided between arrays longer than 10 repeat units (*dark shading*) and 10 units or shorter (*light shading*).

that slippage of 9-base pair and larger units escapes these error-checking processes. Alternatively, the high frequency of polyglutamine repeats in *Drosophila* may indicate that the species has evolved a special use for polyglutamine tracts in proteins. Given the instability of tandem codon repeats, it may be possible to maintain such a large number of repeats only without an unacceptable load of triplet expansion-like disease [which can also be observed in transgenic experiments with *Drosophila* (Bonini 1999)] by selectively eliminating codon structures that are prone to slippage.

A more indirect selective force that could act on the expansion and contraction of polyglutamine repeats is genome size, which has been shown to correlate with the abundance of simple sequences (repeats of short sequence motifs) within a genome (Hancock 1995, 1996). However, there is not a direct relationship between genome size and reiteration of glutamine codons in this study. For example, *S. cerevisiae,* the eukaryotic species that has by far the smallest genome among those studied, appears to have a greater tendency to glutamine codon reiteration than *C. elegans* and *D. melanogaster,* and *C. elegans* in turn contains more codon expansions than *D. melanogaster.* This is probably because these previous analyses of sequence simplicity (Hancock 1995, 1996) were dominated by noncoding regions, which are likely to accumulate repetitive sequences in a more neutral manner than coding sequences.

We have shown that the relative contributions of slippage or slippage-like processes (that is, processes that might give rise to patterns of evolution that could be explained by slippage) and point mutation-like processes differ considerably between diverse species. Slippage appears to have been particularly influential in mammals and almost-absent in *Drosophila.* It should be noted that this conclusion applies to glutamine repeats as analyzed

here, but there is evidence of a contribution of slippage to the wider evolution of the *Drosophila* genome (e.g., Hancock 1996; Hancock et al. 1999). We also see some other patterns, such as higher apparent levels of slippage at CAA in *Arabidopsis* and repetition at the level of three codons in *Drosophila*, which suggest that the underlying mechanisms generating glutamine repeats in proteins may differ between taxonomic groups. Detailed dissection of these processes will require analysis of the evolution of these structures in different taxa, probably considering evolutionary patterns using phylogenetic trees of fairly closely related species (e.g., Hancock and Vogler 2000).

Acknowledgments. We thank the U.K. Medical Research Council for their support. M.M.A. was supported by a postdoctoral fellowship from the Ministerio de Educación y Cultura, Spanish Government.

References

Agianian B, Leonard K, Bonte E, Van der Zandt H, Becker PB, Tucker PA (1999) The glutamine-rich domain of the Drosophila GAGA factor is necessary for amyloid fibre formation in vitro, but not for chromatin remodeling. J Mol Biol 285:527–544

Albà MM, Santibáñez-Koref MF, Hancock JM (1999a) 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

Albà MM, Santibáñez-Koref MF, Hancock JM (1999b) Conservation of polyglutamine tract size between mice and humans depends on codon interruption. Mol Biol Evol 16:1641–1644

Bernardi G (2000) Isochores and the evolutionary genomics of vertebrates. Gene 241:3–17

Bhandari R, Brahmachari SK (1995) Analysis of CAG/CTG triplet repeats in the human genome: Implication in transcription factor gene regulation. J Biosci 20:613–627

Bonini NM (1999) A genetic model for human polyglutamine-repeat disease in *Drosophila melanogaster*. Philos Trans R Soc Lond [Biol] 354:1057–1060

Brock GJ, Anderson NH, Monckton DG (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

Choong CS, Kemppainen JA, Wilson EM (1998) Evolution of the primate androgen receptor: A structural basis for disease. J Mol Evol 47:334–342

Djian P, Hancock JM, Chana HS (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

Eisen JA (1999) Mechanistic basis for microsatellite instability. In: Goldstein DB, Schlötterer C (eds) Microsatellites evolution and applications. Oxford University Press, Oxford, pp 34–48

GCG (1997) Wisconsin package version 9.1. Genetics Computer Group (GCG), Madison, WI

Gerber HP, Seipel K, Georgiev O, Hofferer M, Hug M, Rusconi S, Schaffner W (1994) Transcriptional activation modulated by homopolymeric glutamine and proline stretches. Science 263:808–811

Green H, Wang N (1994) Codon reiteration and the evolution of proteins. Proc Natl Acad Sci USA 91:4298–4302

Hancock JM (1993) Evolution of sequence repetition and gene duplications in the TATA-binding protein TBP (TFIID). Nucleic Acids Res 21:2823–2830

Hancock JM (1995) The contribution of slippage-like processes to genome evolution. J Mol Evol 41:1038–1047

Hancock JM (1996) Simple sequences and the expanding genome. Bioessays 18:421–425

Hancock JM, Vogler AP (2000) How slippage-derived sequences are incorporated into rRNA variable-region secondary structure: Implications for phylogeny reconstruction. Mol Phylogenet Evol 14: 366–374

Hancock JM, Shaw PJ, Bonneton F, Dover GA (1999) High sequence turnover in the regulatory regions of the developmental gene hunchback in insects. Mol Biol Evol 16:253–265

Jones CW, Kafatos FC (1982) Accepted mutations in a gene family: Evolutionary diversification of duplicated DNA. J Mol Evol 19: 87–103

Karlin S, Burge C (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, Trifiro MA, Pinsky L (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

Kruglyak S, Durrett RT, Schug MD, Aquadro CF (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

Levinson G, Gutman GA (1987) Slipped-strand mispairing: A major mechanism for DNA sequence evolution. Mol Biol Evol 4:203–221

Limprasert P, Nouri N, Heyman RA, Nopparatana C, Kamonsilp M, Deininger PL, Keats BJ (1996) Analysis of CAG repeat of the Machado-Joseph gene in human, chimpanzee and monkey populations: A variant nucleotide is associated with the number of CAG repeats. Hum Mol Genet 5:207–213

Limprasert P, Nouri N, Nopparatana C, Deininger PL, Keats BJ (1997) Comparative studies of the CAG repeats in the spinocerebellar ataxia type 1 (SCA1) gene. Am J Med Genet 74:488–493

Margolis RL, Abraham MR, Gatchell SB, Li SH, Kidwai AS, Breschel TS, Stine OC, Callahan C, McInnis MG, Ross CA (1997) cDNAs with long CAG trinucleotide repeats from human brain. Hum Genet 100:114–122

Mitchell PJ, Tjian R (1989) Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. Science 245: 371–378

Nakachi Y, Hayakawa T, Oota H, Sumiyama K, Wang L, Ueda S (1997) Nucleotide compositional constraints on genomes generate alanine-, glycine-, and proline-rich structures in transcription factors. Mol Biol Evol 14:1042–1049

Nakamura Y, Gojobori T, Ikemura T (2000) Codon usage tabulated from international DNA sequence databases: status for the year 2000. Nucleic Acids Res 25:244–245

Ohshima K, Wells RD (1997) Hairpin formation during DNA synthesis primer realignment in vitro in triplet repeat sequences from human hereditary disease genes. J Biol Chem 272:16798–16806

Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 85:2444–2448

Pecheux C, Gall AL, Kaplan JC, Dode C (1996) Sequence analysis of the CAG triplet repeats region in the Huntington disease gene (IT15) in several mammalian species. Ann Genet 39:81–86

Perutz MF, Johnson T, Suzuki M, Finch JT (1994) Glutamine repeats as polar zippers: Their possible role in inherited neurodegenerative diseases. Proc Natl Acad Sci USA 91:5355–5358

Reddy PS, Housman DE (1997) The complex pathology of trinucleotide repeats. Curr Opin Cell Biol 9:364–372

Richard GF, Dujon B (1997) Trinucleotide repeats in yeast. Res Microbiol 148:731–744

Rolfsmeier ML, Lahue RS (2000) Stabilizing effects of interruptions on trinucleotide repeat expansions in *Saccharomyces cerevisiae.* Mol Cell Biol 20:173–180

Rubinsztein DC, Amos W, Leggo J, Goodburn S, Ramesar RS, Old J, Bontrop R, McMahon R, Barton DE, Ferguson-Smith MA (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 DC, Leggo J, Coetzee GA, Irvine RA, Ferguson-Smith MA (1995) 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

Schug MD, Mackay TFC, Aquadro CF (1997) Low mutation rates of microsatellite loci in *Drosophila melanogaster*. Nat Genet 15:99–102

Seznec H, Lia-Baldini AS, Duros C, Fouquet C, Lacroix C, Hofmann-Radvanyi H, Junien C, Gourdon G (2000) Transgenic mice carrying large human genomic sequences with expanded CTG repeat mimic closely the DM CTG repeat intergenerational and somatic instability. Hum Mol Genet 9:1185–1194

Sia EA, Kokoska RJ, Dominska M, Greenwell P, Petes TD (1997) Microsatellite instability in yeast: Dependence on repeat unit size and DNA mismatch repair genes. Mol Cell Biol 17:2851–2858

Stallings RL (1994) Distribution of trinucleotide microsatellites in different categories of mammalian genomic sequence: Implications for human genetic diseases. Genomics 21:116–121

Wharton KA, Yedvobnick B, Finnerty VG, Artavanis-Tsakonas S (1985) opa: A novel family of transcribed repeats shared by the Notch locus and other developmentally regulated loci in *D. melanogaster*. Cell 40:55–62

Wilkins RC, Lis JT (1999) DNA distortion and multimerization: Novel functions of the glutamine-rich domain of GAGA factor. J Mol Biol 285:515–525

Xiao H, Jeang KT (1998) Glutamine-rich domains activate transcription in yeast *Saccharomyces cerevisiae*. J Biol Chem 273:22873–22876