J Mol Evol (2004) 59:329–339
DOI: 10.1007/s00239-004-2628-5

JOURNAL OF MOLECULAR EVOLUTION

Comparative Genetics of Functional Trinucleotide Tandem Repeats in Humans and Apes

Aida M. Andrés,[1] Marta Soldevila,[1] Oscar Lao,[1] Víctor Volpini,[2] Naruya Saitou,[3] Howard T Jacobs,[4] Ikuo Hayasaka,[5] Francesc Calafell,[1] Jaume Bertranpetit[1]

[1] Unitat de Biologia Evolutiva, Facultat de Ciències de la Salut i de la Vida, Universitat Pompeu Fabra, Barcelona, Spain
[2] Molecular Genetics Department, Cancer Research Institute Barcelona, Spain
[3] Division of Population Genetics, National Institute of Genetics, Mishima, Japan
[4] Institute of Medical Technology and Tampere University Hospital, University of Tampere, Tampere, Finland
[5] Sanwa Kagaku, Kenkyusho Kumamoto Primate Park, Misumi, Japan

Received: 28 November 2003 / Accepted: 21 January 2004

Abstract. Several human neurodegenerative disorders are caused by the expansion of polymorphic trinucleotide repeat regions. Many of these loci are functional short tandem repeats (STRs) located in brain-expressed genes, and their study is thus relevant from both a medical and an evolutionary point of view. The aims of our study are to infer the comparative pattern of variation and evolution of this set of loci in order to show species-specific features in this group of STRs and on their potential for expansion (therefore, an insight into evolutionary medicine) and to unravel whether any human-specific feature may be identified in brain-expressed genes involved in human disease. We analyzed the variability of the normal range of seven expanding STR CAG/CTG loci (SCA1, SCA2, SCA3-MJD, SCA6, SCA8, SCA12, and DRPLA) and two nonexpanding polymorphic CAG loci (KCNN3 and NCOA3) in humans, chimpanzees, gorillas, and orangutans. The study showed a general conservation of the repetitive tract and of the polymorphism in the four species and high heterogeneity among loci distributions. Humans present slightly larger alleles than the rest of species but a more relevant difference appears in variability levels: Humans are the species with the largest variance, although only for the expanding loci, suggesting a relationship between variability levels and expansion potential. The sequence analysis shows high levels of sequence conservation among species, a lack of correspondence between interruption patterns and variability levels, and signs of conservative selective pressure for some of the STR loci. Only two loci (SCA1 and SCA8) show a human specific distribution, with larger alleles than the rest of species. This could account, at the same time, for a human-specific trait and a predisposition to disease through expansion.

Key words: Spinocerebellar ataxia — Trinucleotide repeat expansion — Short tandem repeat evolution — Primates

Introduction

The field of comparative genetics has undergone a recent boom due to its power as a tool for the understanding of the evolutionary and functional factors shaping a given genome region and for the search of the genetic basis of species uniqueness. Despite the large effort made to obtain comparative genetic information, little is still known about similarities and differences between the human genome and that of

Correspondence to: Dr. Jaume Bertranpetit; *email:* jaume.bertranpetit@upf.edu

Table 1. Genetic characteristics of the nine loci examined in this study

Locus	MIM	Chromosome	Repeat	Region	Expansion
SCA1	601556	6p23	CAG	Coding	Yes
SCA2	601517	12q24	CAG	Coding	Yes
SCA3	607047	14q24-q32	CAG	Coding	Yes
SCA6	601011	19p13	CAG	Coding	Yes
SCA8	603680	13q21	CTG	Untranslated	Yes
SCA12	604326	5q31-33	CAG	5′ UTR	Yes
DRPLA	125370	12p13	CAG	Coding	Yes
KCNN3	602983	20q12	CAG	Coding	No
NCOA3	601937	1q21	CAG	Coding	No

Note. Chromosome, gene chromosomal localization; repeat, repeat unit; region, location of the repetitive region within the gene; expansion, expanding (yes) or nonexpanding (no) loci.

our closest phylogenetic relatives, the apes. In recent years, several human-specific genetic traits have been detected by a comparative analysis of human and primate species, mostly from the analysis of genetic regions in chimpanzees (for reviews, see Gagneux and Varki 2001; Hacia 2001). Comparative information has mainly been obtained for noncoding regions, and such data have been the key to the understanding of hominoid phylogeny (Chen and Li 2001; see Ruvolo 1997 for a previous review). Comparative analyses of functional regions usually aim to understand the functional constraints on a particular genetic region, including purifying, positive, and balancing selection. In some cases, analysis may help to explain the appearance of a new genetic variant in a particular species, such as lysozyme enzymes in primates (Messier and Stewart 1997) or in the FOXP2 gene in humans (Lai et al. 2001; Enard et al. 2002).

We focused our study on the comparative analysis of a special group of genetic elements: functional CAG/CTG repetitive tracts. These are functional CAG/CTG short tandem repeats (STRs), mostly coding for polyglutamine tracts; they are found in genes that are highly expressed in the brain and their expansion in repeat number causes neurodegenerative disorders. Many of the genes in this study cause spinocerebellar ataxia (SCA) and are named, accordingly, SCA*n* loci, where *n* denotes a locus defining number.

These type of loci share the mutation dynamics of the rest of STRs of the genome, that is, they mutate by adding or subtracting one (or rarely more than one) repeat unit, but differ by expanding into abnormally long alleles that produce ataxia, dystrophy or similar diseases. This group of genetic diseases has not been detected in nonhuman species, a fact that could just be attributed to the greater knowledge of human disease but might also reflect a unique human pathogenic trend. Beyond the mere description, the comparison of the patterns of variability in the normal range of humans and apes can allow the testing of hypotheses on the causes of expansion and disease. This study can therefore be viewed in the context of evolutionary medicine, where the comprehension of the natural history of a disease may lie in the particular characteristics (either of the locus or of the species) of disease predisposition.

The variability of expanding loci has been studied in humans in order to determine population-specific disease risk factors and to understand the evolutionary forces shaping this variability (Watkins et al. 1995; Jodice et al. 1997; Andrés et al. 2003). On the other hand, interspecific comparative studies on these loci have focused on allele length comparison between humans and other primate species and have dealt mostly with a few species used as references in a single locus approach. The special interest in allele length differences among species is based on the observation that long alleles have an increased mutation rate and higher probability of very long leaps (Webster et al. 2002) and the possibility of expansion into the pathogenic range (as shown [Fu et al. 1991; Nolin et al. 2003] for fragile X).

We have analyzed, in four species (human, chimpanzee, gorilla, and orangutan), nine STR loci including those present on the SCA1 (spinocerebellar ataxia 1 locus), SCA2, SCA3 (or Machado–Joseph disease locus), SCA6, SCA8, SCA12, DRPLA (dentatorubral–pallidoluysian atrophy), KCNN3 (potassium intermediate/small conductance calcium-activated channel, subfamily N, member 3), and NCOA3 (nuclear receptor coactivator 3) genes (Table 1). They all are functional, have different genomic locations and functions, and share a high central nervous system expression and the presence of a CAG/CTG repetitive tract. Seven are expanding disease-related loci, while the remaining two are coding but do not seem to expand into pathogenic alleles. Variation at KCNN3, other than expansion, has been proposed as being associated with mental diseases and ataxia (Dror et al. 1999; Figueroa et al. 2001), although as a predisposing factor rather than as a single direct cause.

A strong correlation between expansion and variability in the normal range has been demonstrated by the observation that expanding STRs are the most variable group of STRs of the human genome

(Chakraborty et al. 1997; Jodice et al. 1997; Deka et al. 1999). As these expansions have not been detected in apes, our aim is to determine whether these loci show similar levels of variability in apes to those observed in humans (and therefore diversity and expansion potential would not be directly related) or whether the high levels of polymorphism are exclusive to the species for which the expanding disease has been detected.

We have also determined whether shared factors among loci (such as the existence of a coding poly [CAG] tract) or locus-specific ones (such as differential mutation patterns or selective events) led to the diversity of observed allele distribution of expanding loci.

Ascertainment bias may be a relevant problem when trying to infer general patterns from a group of loci selected in one of the species or populations compared. Nevertheless, this problem does not affect our study, as we are interested in determining the evolution of this specific group of loci (where the observed tendencies are clear in terms of expansion and disease) but do not try to generalize our observations to infer traits for the rest of STRs, which would produce a strong ascertainment bias.

Materials and Methods

Allele Typing

Twenty common chimpanzees (*Pan troglodytes* subspecies *troglodytes* [one individual] and *verus*), 13 gorillas (*Gorilla gorilla* subspecies *gorilla*), and 4 to 6 orangutans (*Pongo pygmaeus* subspecies *abelii* [two individuals] or unknown) were typed for the number of repeats at the nine loci shown in Table 1. Data for SCA1 and SCA3 in chimpanzees (16 individuals) were obtained from the literature (Limprasert et al. 1996, 1997). In order to avoid sampling a single primate population, we obtained the samples from very different sources: Coriell Cell Repositories (USA), European Collection Cell Cultures (UK), Barcelona Zoo (Spain), Kumamoto Primate Park (Japan), Institute of Zoology, London (UK), and Dr. Takafumi Ishida, Tokyo (Japan). Although it is not possible to assume that this sample set is representative of genetic variation of all apes, care was taken to try to obtain a diverse sample set, whose heterogeneous origin reduces the limitations of studying a small population sample. Subspecies of ape specimens were determined by amplification by PCR of both hypervariable regions of the mtDNA D-loop region and direct sequencing with internal primers. Subspecies identification may not be crucial for our study, however, as variation at nuclear loci does not cluster by subspecies, at least for chimpanzees (Kaessmann et al. 1999).

All DNA samples were amplified for the region containing the repetitive element by PCR, with primers and conditions previously described for human analysis (Table A1). In the KCNN3 gene, which contains two CAG repetitive segments, the most variable repetitive tract (nucleotides 513 to 569 in NCB1 Reference Sequence NM_002249.3) was amplified in a fragment that did not include the other repetitive tract. Lengths of the amplified fragments were subsequently typed with GeneScan software version 3.7 (Applied Biosystems) after electrophoresis on 6% denaturing gel performed on an ABI Prism377 automatic sequencer (Applied Biosystems). A human sample of known genotype, previously sequenced and typed, was used as a standard size control for all GeneScan runs.

At least one individual per species and locus was directly sequenced to verify the correspondence between amplified fragment length and number of repeats. Sequences were determined using BigDyes sequencing kits, versions 2.0 and 3.0 (Applied Biosystems), on automatic sequencers ABIPrism 377 and 3100 (Applied Biosystems). The only exception is DRPLA, for which ape sequences from GenBank were used (accession number AJ133270 [*Pan paniscus*], AJ133271 [*Gorilla gorilla*], AJ133272 [*Pongo pygmaeus*]). Sequences were assembled and analyzed with the SeqmanII program (Lasergene 1999 package; DNASTAR, Inc.). Homology between human and mouse sequences was obtained by BLAST, additional alignment of sequences (Seqman II and Clustal programs), and manual determination of the conserved regions.

Allele Frequencies and Statistical Analysis

The correspondence between allele size and number of repeats for every species allowed the estimation of allele frequencies by direct allele counting. Human distributions for healthy individuals result from pooling African, European, Indian, and East Asian samples for all loci; the distributions were obtained from the literature or the typing of healthy individuals when necessary (Andrés et al. 2003). As they come from normal individuals, we do not expect to find a significant proportion of premutated alleles (those with an intermediate allele length, i.e., between normal and pathogenic, and which have high expansion probability). Allele frequencies were averaged across populations without weighing by sample size, as the number of chromosomes varied greatly among populations. All subsequent analyses were performed with the pooled human distribution.

The different parameters of allele size distributions for every locus and species were determined. Mean, standard deviation, variance, and variation coefficient of repeat number were calculated with the SPSS statistical package. Expected heterozygosity was calculated with the Arlequin 2.000 package (Schneider et al. 2000). In order to analyze the possibility that variance was influenced by differences in sample size, we computed it on an increasing sample size, for every locus and species, with a program that takes a pseudosample of 8–40 chromosomes (the range in size between our smallest and our largest ape samples) from the original distribution and calculates its variance. After 1000 random extractions for each pseudosample size, the average variance for every pseudosample size is plotted in a graph that shows how variance relates to pseudosample size.

To determine whether variances observed in species with a smaller sample size could be obtained from a hypothetical human sample of eight chromosomes (our smallest sample size, that of orangutans), we performed a second resampling experiment. In this case, a pseudosample of eight chromosomes was obtained for every human distribution and its variance was calculated; after 10,000 sample extractions, we obtained a distribution of variances from pseudosamples of eight human chromosomes. To determine the significance of our results, the variance was compared to the 95% confidence interval of the pseudosample variance distribution.

The four species were compared in terms of mean and variance of repeat number. Two tests were performed for mean repeat number with permutation tests: the comparison between the four species for every individual locus (considering every species as a different category) and the comparison between humans and the rest of the species for every locus (considering "apes" a category, which included all ape species, and comparing it with the "human" category). The permutation test was performed as follows: Individual chromosomes were randomly shuffled between classes (species or species groups), maintaining the original sample sizes. For every permutated data set, the difference in the average number of repeats between the two classes was computed. This process was

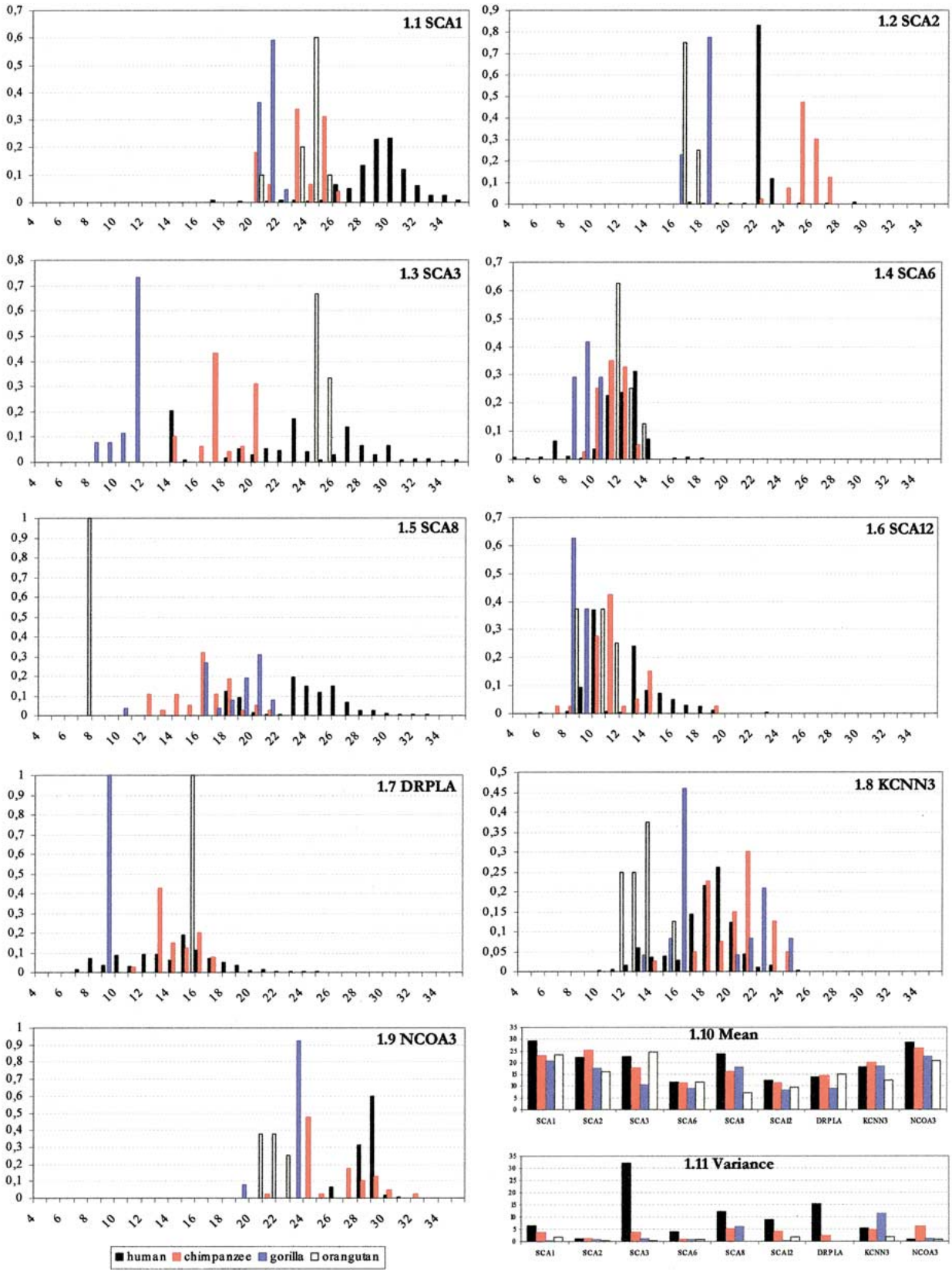

Fig. 1. Graphical representations of individual allele size distributions and average length and variability parameters for species and loci. **1.1–1.9** Allele length distributions in the four species. The distributions are the result of typing, for repeat number, chimpanzee (in red), gorilla (blue), and orangutan (yellow) chromosomes, and pooling human distributions from African, European, Indian, and East Asian origin (black). **1.10** Mean allele size for species and locus. **1.11** Variance of repeat number for species and locus.

repeated 1000 times. The test is significant if the probability of obtaining a difference in average repeat number in the permutations as large as in the observed data set, in a one-tailed test, is <0.05 for the human–ape comparison or <0.0056 (after Bonferroni correction) for the four-species comparison.

Variance comparison among the four species, for every locus, was performed with Scheffé–Box (log ANOVA) test for homogeneity of variances (Sokal and Rohlf 1995, p. 397); as this test is not available in statistical packages, an ad hoc program has been written and is available on request (oscar.lao@upf.edu). In order to obtain a single overall significance value for all loci, the individual Scheffé–Box p values for every locus were combined with Fisher's test for probability combination (Sokal and Rohlf 1995, p. 795). When a species showed a single allele for a given locus the test was performed using the rest of the species.

Results

Data on nine CAG repeat loci, whose characteristics are shown in Table 1, were obtained for humans (of African, European, Indian, and East Asian origin), for 20 chimpanzees, 13 gorillas, and 4 to 6 orangutans; all chromosomes analyzed were in the "normal," nonexpanded range. Allele size distributions are plotted by locus in Fig. 1, and statistical parameters of the distributions and expected heterozygosity are shown in Table 2. The amplified fragment was sequenced in at least one individual for species and locus; sequences of all repetitive regions are shown in Fig. 2, and their accession numbers are available in Table A2, including compared mouse loci. Comparative locus analysis was carried out on different levels, searching for general species trends and for locus-specific trends.

General species-specific trends can be detected by comparing the parameters of the distribution for every species over all loci (Table 3). Humans show a higher mean number of repeats than the other species, reaching statistical significance (permutation test humans vs. apes, $p < 0.001$; see Materials and Methods for the groups considered), showing that a trend exists in allele length among the different species. Nevertheless, the trend is not followed by all loci, as discussed below.

Variance and coefficient of variation of repeat number show a decreasing trend from humans to orangutans, as shown in Table 3. Statistically significant differences in variance exist between species, with all individual loci showing significant differences among species ($p < 0.05$) and the combined p value also being statistically significant. Interestingly, for the seven expanding loci humans presented the largest variance.

The small sample size in the ape species may bias the estimation of the dispersion parameters. As stated in Materials and Methods, the samples for nonhuman species are of very heterogeneous origin, which reduces the possibility of underestimating variance by sampling from a single, localized population or inbred zoo collection. Furthermore, to explore to what extent human variance is larger than that of the rest of the species as a consequence of its larger sample size, we performed two independent tests.

First, we studied whether variance increased with sample size by a permutation test on pseudosamples of 8–40 chromosomes for the four species in every locus. Results showed that reduction of sample size does not lead to a reduction of variance (data not shown) and that the resampling average value remains within the range of the values obtained from the original distribution. Therefore, as expected, a reduced sample does not determine lower variance

Table 2. Sample size, statistical parameters, and expected heterozygosity for each species and locus

Locus	Species	*N*	Mean	SD	Var.	CV	*H*
SCA1	Human	1098	29.29	2.55	6.52	8.72	0.850
SCA1	Chimpanzee	32	23.09	1.91	3.64	8.32	0.764
SCA1	Gorilla	22	20.68	0.57	0.32	2.78	0.541
SCA1	Orangutan	10	23.50	1.35	1.83	5.91	0.644
SCA2	Human	1214	22.15	1.10	1.22	4.98	0.291
SCA2	Chimpanzee	40	25.40	0.98	0.96	3.89	0.680
SCA2	Gorilla	22	17.55	0.86	0.74	4.94	0.368
SCA2	Orangutan	08	16.25	0.46	0.21	2.94	0.429
SCA3	Human	1164	22.72	5.68	32.29	25.01	0.890
SCA3	Chimpanzee	32	17.75	1.88	3.55	10.70	0.716
SCA3	Gorilla	26	10.50	0.95	0.90	9.12	0.459
SCA3	Orangutan	12	24.33	0.49	0.24	2.07	0.485
SCA6	Human	1130	11.76	2.02	4.07	17.17	0.786
SCA6	Chimpanzee	40	11.13	0.94	0.88	8.49	0.724
SCA6	Gorilla	24	9.00	0.78	0.61	8.76	0.685
SCA6	Orangutan	8	11.50	0.76	0.57	6.78	0.607
SCA8	Human	1824	23.59	3.51	12.29	14.87	0.876
SCA8	Chimpanzee	38	16.13	2.26	5.09	14.08	0.848
SCA8	Gorilla	26	18.15	2.43	5.90	13.50	0.812
SCA8	Orangutan	08	7.00	0.00	0.00	0.00	0.000
SCA12	Human	3336	12.27	3.01	9.08	24.56	0.781
SCA12	Chimpanzee	40	11.33	2.02	4.07	17.93	0.735
SCA12	Gorilla	24	8.38	0.49	0.24	5.97	0.489
SCA12	Orangutan	8	9.50	1.31	1.71	14.21	0.750
DRPLA	Human	809	14.07	3.94	15.53	28.02	0.907
DRPLA	Chimpanzee	40	14.25	1.50	2.24	10.58	0.754
DRPLA	Gorilla	26	9.00	0.00	0.00	0.00	0.000
DRPLA	Orangutan	8	15.00	0.00	0.00	0.00	0.000
KCNN3	Human	968	17.88	2.31	5.32	12.90	0.840
KCNN3	Chimpanzee	40	20.05	2.16	4.66	10.84	0.831
KCNN3	Gorilla	24	18.29	3.36	11.26	18.54	0.754
KCNN3	Orangutan	8	12.50	1.31	1.71	10.80	0.821
NCOA3	Human	880	28.50	0.90	0.80	3.15	0.540
NCOA3	Chimpanzee	40	26.00	2.47	6.10	9.56	0.732
NCOA3	Gorilla	26	22.69	1.09	1.18	4.84	0.148
NCOA3	Orangutan	8	20.88	0.83	0.70	4.12	0.750

Note. N, sample size; mean, mean number of repeats; SD, standard deviation; var., variance; CV, coefficient of variation; H, expected heterozygosity. Human parameters were calculated from the pooled distribution (see Materials and Methods).

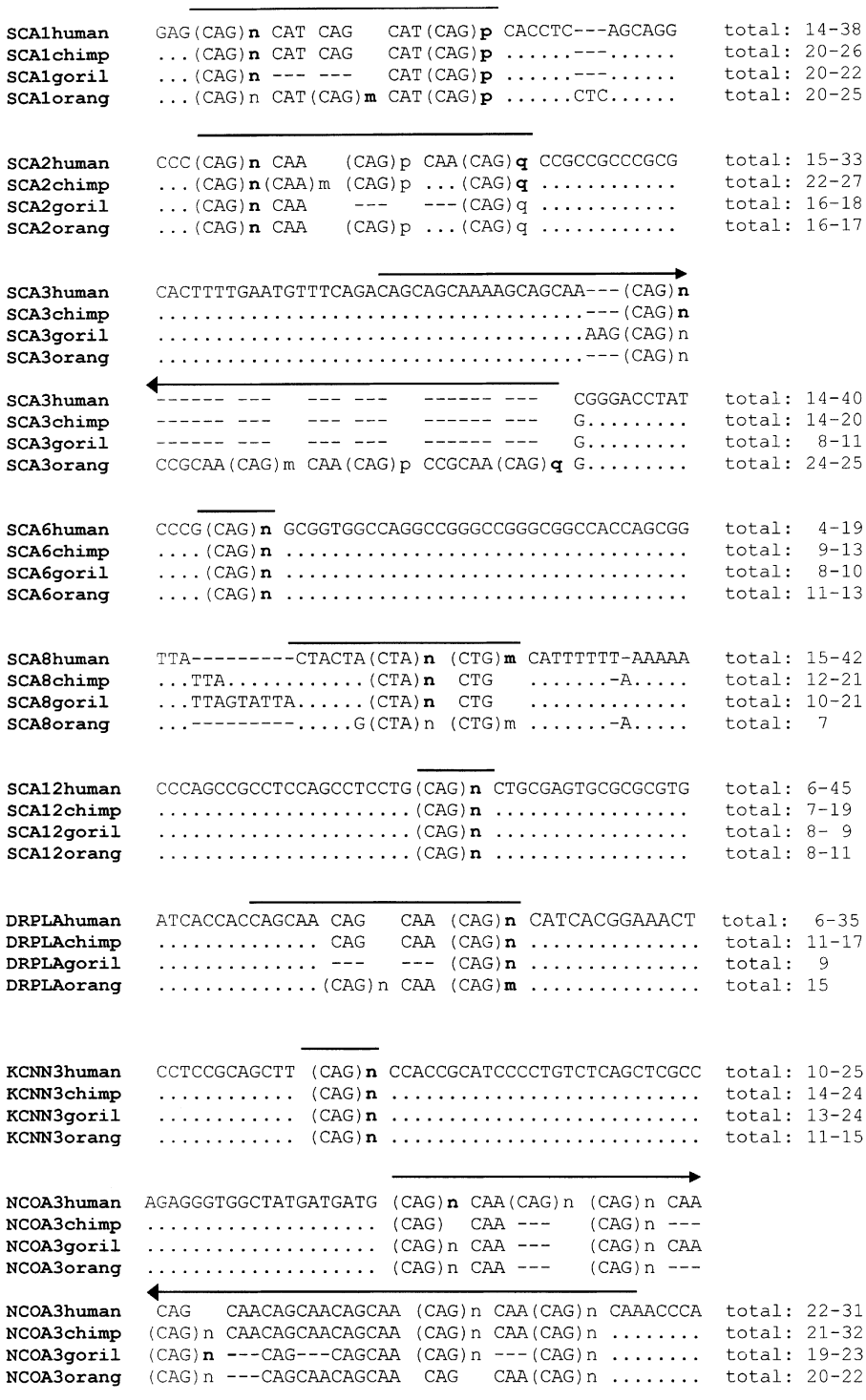

Fig. 2. Sequence of the repetitive region for the expanding loci in the four species. The line over the sequences marks the STR sequence region, and the total length of the overlined sequence region is shown. Repetitive segments are indicated without detailing the exact number of repeats, and segments containing six or more repeat units, with higher probabilities of slippage events than shorter segments, are marked in bold face.

value. A second resampling test was performed, in which we tested whether variances similar to those obtained for apes could be obtained from pseudo-samples of eight human chromosomes from our human distributions (see Materials and Methods). Of the seven loci with larger variance in humans than in the rest of the species, four loci (SCA3, SCA8, SCA12, and DRPLA) showed large and statistically significant differences ($p < 0.05$) between humans and orangutans (the species with the lowest sample size), and two of them (SCA3 and DRPLA) showed significantly larger variances in humans than in any other species. This analysis suggests that beyond some sample size influence, our results cannot be exclusively explained for differences in sample size, and variance divergence for this group of loci seems to be a species characteristic.

Heterozygosity is high and very similar between humans and chimpanzees and low in gorillas and orangutans (Table 3). This observation is due to the

Table 3. Parameters of allele size distributions for species: Pooled sample size, statistical parameters, and expected heterozygosity for each species in the different loci

Species	*N*	Mean	SD	Var.	CV	*H*
Human	12423	20.25	2.78	9.68	15.49	0.751
Chimpanzee	342	18.35	1.79	3.47	10.49	0.754
Gorilla	220	14.92	1.17	2.35	7.60	0.473
Orangutan	78	15.61	0.72	0.78	5.20	0.498

Note. Data for all loci were pooled for each species. Abbreviations are the same as those used in Table 2; see Note there.

large influence of SCA2 and NCOA3, with higher heterozygosities in chimpanzees than in humans; without these two loci, mean heterozygosity is higher in humans than in chimpanzees.

In order to determine which loci are mostly responsible for the species trends, mean repeat number (Fig. 1.10) and variance of repeat number (Fig. 1.11) were compared for each species and locus. Humans do not always show larger allele sizes at all loci or at all expanding loci, and only three (SCA1, SCA8, and NCOA3) show statistically significant larger mean allele length in humans than in any other species. On the other hand, variance is higher in humans than in the rest of species for all expanding loci but not for the two nonexpanding loci. Nonetheless, the low number of nonexpanding loci analyzed prevents us from generalizing this interesting difference in variance.

The individual comparison among loci distributions shows a strong heterogeneity in allele size distribution (Fig. 1), and sequences of the STR alleles over the four species (Fig. 2) illustrate that repeat regions are very complex and that different loci and lineages show heterogeneous amounts and patterns of divergence. The sequence distribution comparison clearly shows a lack of general patterns (in STR sequence and interrupting complexity) that could explain the observed allele size and variability of expanding loci.

Interestingly, when comparing human and mouse, only very short tracts were detected for almost all mouse loci, from the complete absence of the tract (in SCA6, SCA8, and SCA12 the repetitive region could not be detected) to very short CAG/CAA repetitive regions (four CAG repeats in SCA1, two in SCA2, five in SCA3, four in DRPLA, and one in NCOA3). Only KCNN3 keeps the repetitive tract over phylogenetically distant species, as a repetitive CAG/CAA region exists in the mouse sequence, with the interrupting codons CAA (Gln) and TCG (Ser), which are absent in primates.

Discussion

Our study shows that the "flexible conservation" that exists in human functional CAG/CTG tracts (high polymorphism in number of repeats combined with sequence conservation) is also found in apes. The conservation of the variable tracts in all species strongly supports the presence of the repeat regions in the common ancestor of humans and apes in all nine loci studied, and the presence of polymorphism in almost all species suggests the existence of ancestral polymorphism.

Previous studies comparing human and mouse CAG/CAA tracts found a relationship between interrupting levels and conservation between very distant species (Albà et al. 1999), suggesting that older tracts would be more frequently interrupted by non-CAG codons. We failed to find this relationship in our set of repetitive tracts, as the most conserved locus (KCNN3, with a long repetitive region in both humans and mice) is totally uninterrupted in primates, and many tracts showing low conservation between distant species are profoundly interrupted in primates.

Species-Specific Characteristics

The combined analysis of all loci points to species-specific trends. In allele length comparison, larger alleles in humans than in other species were reported for nonfunctional STRs (Rubinsztein et al. 1995a; Crouau-Roy et al. 1996; Cooper et al. 1998), a controversial conclusion (Ellegren et al. 1995, 1997). Previous studies found longer alleles in humans than in other species for SCA1, SCA3, AR, and HD CAG repeats and in the FA (GAA) locus (Rubinsztein et al. 1995b; Djian et al. 1996; Limprasert et al. 1997; Choong et al. 1998; Gonzalez-Cabo et al. 1999; Justice et al. 2001). This suggests a general increase in the number of repeats from monkeys to apes and humans for expanding loci (although similar allele length was found by Limprasert et al. [1996] in the SCA3 locus). On the other hand, a study on functional nonexpanding CAG STR shows shorter alleles in humans than in apes (Saleem et al. 2001). Therefore, a clear picture of the human specific characteristics in functional STRs had not emerged beyond single locus comparisons.

The data presented in this paper for nine functional trinucleotides show significantly higher number of repeats in humans, a trend that is unique to this lineage and shows that, beyond locus heterogeneity, a specificity in allele length exists in this set of STRs. The trend is not present in all loci and is mainly (but not exclusively) due to SCA1, SCA8, and NCOA3.

In our set of samples and loci, we found differences in variance among the four species, with the highest values in humans. In the genomic regions studied so far, DNA sequence diversity is lower in humans than in chimpanzees and other apes (Crouau-Roy et al. 1996; Kaessmann et al. 2001; Noda et al. 2001), possibly due to a demographic bottleneck in the human lineage (Jorde et al. 2000). Surprisingly, the present results show that humans are more diverse

(measured as variance of allele distribution) than any of the other species studied for an ample set of expanding functional trinucleotide tandem repeats. No demographic factor (which would affect the whole genome) or ascertainment bias effect for the selection of the STRs in humans could explain this finding, which is not a general STR trend, but it is exclusive of functional STRs that can expand and produce disease. This is not found in noncoding STRs (and previous results [Crouau-Roy et al. 1996; Garza et al. 1995; Wise et al. 1997] are not concordant) or in other similar STRs (such as the nonexpanding KCNN3 and NCOA3, analyzed in this study).

The possible sampling error in apes has been reduced by choosing individuals of different origins as much as possible (see Materials and Methods), and the effects of different sample size for the different species have been proven to be small through resampling procedures (see Results). Moreover, variability levels may be investigated to test whether our primate samples present lower variability levels than humans at other genetic loci. Seventeen of the chimpanzees typed in this study were previously analyzed for 16S rRNA (Noda et al. 2001); the subset of individuals analyzed in both studies show variability levels ($\pi = 0.0014 \pm 0.0003$) comparable with those existing for the 16S rRNA of the whole human species ($\pi = 0.0016 \pm 0.0004$) (Ingman et al. 2000). These results show that a part of our sample of chimpanzees is as variable as humans at a global scale, and thus the expected diversity of chimpanzees in CAG repeats should be higher than humans. Therefore, our results are not the consequence of a small sample size or the selected sample of individuals, and locus-specific factors (related to CAG loci) acting in different ways in different species are needed to explain the observation that the loci that can expand in humans are more variable in humans than in any other species.

Human expanding STRs have previously been shown to be more variable than other di-, tri-, or tetranucleotides in the human genome (Chakraborty et al. 1997; Jodice et al. 1997; Deka et al. 1999), suggesting a relationship among variability, expansion, and disease. Moreover, our results show that these loci are more variable in humans than in any ape, suggesting that this pattern could be related to a human specifically capacity for expansion. The relationship suggests that loci with increased variance may be more likely to expand to pathogenic alleles, and thus lead to disease.

Different scenarios would be compatible with the increase in variance in one species: the first is a high mutation rate in the absence of strong selective constraints, which would increase variability of the STR (Di Rienzo et al. 1998), leading to new alleles that, if long enough, would increase slippage probability to expanded alleles. Humans do not have mean larger alleles, and therefore allele length differences do not seem to explain the observation. Nevertheless, we observed a statistically significant relationship between the longest allele and variance, even correcting for influence of mean allele size ($r = 0.6552$, $p < 0.0005$). The observed relationship between variance and longest allele can be explained for the nature of variance calculation (as alleles far away from the mean allele size will strongly influence variance). Thus normal but long alleles might affect the STR dynamics: as mutation rate increases with allele size, very long alleles can contribute to the high mutation rate and high variability levels. But this factor, although important for the production of pathogenic alleles, would have a minor effect on the overall variance, as they are rare in the population, and its increased mutation rate will probably not be large enough to explain the huge differences observed in variance for the whole distribution. When computing variances without the very long alleles (eliminating the five longer alleles in a range of less than 40 repeats), variance decreases only 20% on average, far from the differences with other species, showing that very long alleles are not the main responsible alleles for the large human variances in expanding loci. No other factors seem to exist to explain differences in mutation rates.

A second possibility, given the functionality of the loci, is the existence of some kind of balancing selection that, as in the case of other human genes like HLA (Hedrick and Thomson 1983; Hughes and Nei 1988) or CCR5 (Bamshad et al. 2002), would maintain the high diversity by favoring the existence of different alleles in the population. This is a truly speculative explanation, as no external evidence exists for the presence of balancing selection acting over these loci, but the fact that they are functional require consideration of a nonneutral explanation for our findings.

A detailed knowledge both of the mutation rate and pattern for each locus and of the functional behavior of the different alleles may be necessary to interpret these findings and to infer whether mutation or selection are the main causes of the increase of variability in humans.

Locus-Specific Characteristics

Differences in a single functional locus among species may influence its function, and these differences are especially interesting as these genes are expressed in brain and their variation outside the normal range has pathogenic effects. These characteristics (functionality and variability) make poly(CAG) regions attractive candidates for having brain-related functions, especially in the human lineage, which shows

many brain-specific phenotypic traits compared to the rest of species.

The analysis of each locus demonstrates the existence of a strong evolutionary heterogeneity, showing that specific evolutionary forces have been acting on each gene, governing its diversification and evolution. No important differences in allele length exist between humans and apes, the only exception being SCA1 and SCA8, the two candidate loci to present functional differences between humans and the other species if the length of the repetitive tract influenced protein function. This possibility has been suggested for SCA1 (Yue et al. 2001) and SCA8 (see Andrés et al. 2003 for a detailed explanation of SCA8 interspecific differences). The variability levels are also highly heterogeneous among loci, ranging from loci with similar diversity in all species (as in SCA6 and SCA12) to those with extreme differences (SCA3 and DRPLA).

As in allele length distributions, there is a strong heterogeneity in the repetitive sequence conservation among loci for the different species: in contrast to the total conservation of the CAG tract in some loci (SCA6, SCA12, and KCNN3), others show important sequence differences among species, both in the presence of interruptions and in the composition of the repetitive segment (as SCA2, SCA3, SCA8, or DRPLA), which results in differences in repeat number. Nevertheless, the complexity in sequence interruption patterns could not be related to the amount of allele size diversity; and any simple pattern of STR mutation depending on the repetitive sequence that could be inferred from the analysis of one functional STRs locus does not seem to apply to others. Heterogeneity is the main rule.

The comparative study of seven expanding and two nonexpanding brain-expressed functional STRs has shown the overall maintenance of polymorphism with a higher mean in humans for some loci and a larger amount of variability in humans for expanding loci. Comparative studies focused on the search of species specificities on functional regions often do not deal with intraspecific variability, trying to find out fixed changes among species. Here we suggest that differences in genetic variability in important brain-expressed genes should also be considered. Species variability levels on phenotypic traits can also be a species-unique characteristic, and the study of diversity in specific genes with important gene function may give a clue toward understanding the species-specific evolution of these genes.

Appendix

Table A1. Primer sequences used in this study

Locus	Primer	Sequence	Ref
SCA1	rep1	AACTGGAAATGTGGACGTACGCATG	Orr et al. 1993
	rep2	CAACATGGGCAGTCTGAG	Orr et al. 1993
	sca1bR	CGTACTGGTTCTGCTGGG	primers designed for this study
	sca1cF *	CACCAGTGCAGTGGCCTCG	primers designed for this study
	sca1cR *	CGGCCGGTGTTCTGCGGAG	primers designed for this study
SCA2	sca2a	GGGCCCCTAACCATGTCG	Pulst et al. 1996
	sca2b	AGGGCTTGCGGACATTGG	Pulst et al. 1996
	sca2bR *	CTTGCGGACATTGGCAGCC	primers designed for this study
	sca2cF *	GCGAGCCGGTGTATGGGCC	primers designed for this study
SCA3	mjd52	CCAGTGACTACTTTGATTCG	Kawaguchi et al. 1994
	mjd70 *	CTTACCTAGATCACTCCCAA	Kawaguchi et al. 1994
	sca3bF *	ACAATGTATTTTCCTTATGAATAG	primers designed for this study
SCA6	s-5f1	CACGTGTCCTATTCCCCTGTGATCC	Zhuchenko et al. 1997
	s-5r1	TGGGTACCTCCGAGGGCCGCTGGTG	Zhuchenko et al. 1997
	sca6bF *	CCTGTGATCCGTAAGGCCGG	primers designed for this study
	sca6cR *	GCCTGGGACCCGCCTCTCC	primers designed for this study
SCA8	sca8-f3 *	TTTGAGAAAGGCTTGTGAGGACTGAGAATG	Koob et al. 1999
	sca8-r4 *	GGTCCTTCATGTTAGAAAACCTGGCT	Koob et al. 1999
SCA12	sca12mF *	CCACTGCAGCAAAGAGCAGC	Andrés et al. 2003
	sca12mR *	TGCAGTGGCGAGATGGCAGG	Andrés et al. 2003
DRPLA	b37r	CACCAGTCTCAAACAATCACCATG	Nagafuchi et al. 1994
	b33f *	CCTCCAGTGGGTGGGGAAATGCTC	Nagafuchi et al. 1994
	drplacF *	CCACCACCTCCTCCCTATGG	primers designed for this study
	drplacR *	GGAGACATGGCGTAAGGGTGTG	primers designed for this study
KCNN3	kcnn3mF *	TCGCTGCAGCCTCAGCCTC	Andrés et al. 2003
	kcnn3mR *	GCCAGGCCCACTTGTAGCTG	Andrés et al. 2003
NCOA3	ncoa3mF *	GAGAGCTGCTAAGTCATCAC	Andrés et al. 2003
	ncoa3mR *	GTTGATATGGAAACTGTTGCG	Andrés et al. 2003

Note. *, primers used for sequencing. Original reference for each primer is given.

Table A2. New nucleotide data accession numbers

Locus	Chimpanzee	Gorilla	Orangutan
SCA1	AY225351	AY225352	AY225353
SCA2	AY225366	AY225367	AY225368
SCA3	AY225369	AY225370	AY225371
SCA6	AY225372	AY225373	AY225374
SCA8	AY225354	AY225355	AY225356
SCA12	AY225357	AY225358	AY225359
KCNN3	AY225360	AY225361	AY225362
NCOA3	AY225363	AY225364	AY225365

Note. Mouse accession numbers: Mouse SCA1[NM_009124], SCA2 [NM_009125], SCA3 [XM_127081], SCA6 [NM_007578], SCA8 [AF252281], DRPLA [NM_007881], KCNN3 [AF357241], NCOA3 [NM_008679].

Acknowledgments. We thank Monica Vallés for her technical support and Anna Pérez-Lezaun for her technical help and feedback. The authors would like to especially thank Arcadi Navarro for reading the manuscript and for his help with statistical comparisons, Michael Greenacre for statistical help, Lynn B. Jorde, Pascal Gagneux, and Mar Albà for reading an early version of the manuscript, and Lynda Vigilant, Pascal Gagneux, and Anne C. Stone for helpful advice on primate subspecies determination. Tanzanian human samples were kindly supplied by Dr. Clara Menéndez from the Unitat d'Epidemiologia i Bioestadística (Hospital Clínic, Barcelona). Primate samples were supplied by the Barcelona Zoo (under the agreement of the Primate DNA Bank with the Pompeu Fabra University), Ikuo Hayasaka, and Dr. Takafumi Ishida of the University of Tokyo (Japan) or obtained from the Institute of Zoology (London), Coriell Cell Repositories (CCR), and European Collection of Cell Cultures (ECACC). This study was supported by the Dirección General de Investigación (Spanish Goverment), Grants SAF 2001-0772 to J.B. and BOS2001-0794 to F.C., and Grants-in-Aids to Priority Area from MEXT, Japan, to N.S. A.M.A. was financially supported by a fellowship from Generalitat de Catalunya, 2000FI 00686.

References

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

Andres AM, Lao O, Soldevila M, Calafell F, Bertranpetit J (2003) Dynamics of CAG repeat loci revealed by the analysis of their variability. Hum Mutat 21:61–70

Bamshad MJ, Mummidi S, González E, Ahuja SS, Dunn DM, Watkins WS, Wooding S, Stone AC, Jorde LB, Weiss RB, Ahuja SK (2002) A strong signature of balancing selection in the 5′ cis-regulatory region of CCR5. Proc Natl Acad Sci USA 99:10539–10544

Chakraborty R, Kimmel M, Stivers DN, Davison LJ, Deka R (1997) Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. Proc Natl Acad Sci USA 94:1041–1046

Chen FC, Li WH (2001) Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am J Hum Genet 68:444–456

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

Cooper G, Rubinsztein DC, Amos W (1998) Ascertainment bias cannot entirely account for human microsatellites being longer than their chimpanzee homologues. Hum Mol Genet 7:1425–1429

Crouau-Roy B, Service S, Slatkin M, Freimer N (1996b) A fine-scale comparison of the human and chimpanzee genomes: linkage, linkage disequilibrium and sequence analysis. Hum Mol Genet 5:1131–1137

Deka R, Guangyun S, Smelser D, Zhong Y, Kimmel M, Chakraborty R (1999) Rate and directionality of mutations and effects of allele size constraints at anonymous, gene-associated, and disease-causing trinucleotide loci. Mol Biol Evol 16:1166–1177

Di Rienzo A, Donnelly P, Toomajian C, Sisk B, Hill A, Petzl-Erler ML, Haines GK, Barch DH (1998) Heterogeneity of microsatellite mutations within and between loci, and implications for human demographic histories. Genetics 148:1269–1284

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

Dror V, Shamir E, Ghanshani S, Kimhi R, Swartz M, Barak Y, Weizman R, Avivi L, Litmanovitch T, Fantino E, Kalman K, Jones EG, Chandy KG, Gargus JJ, Gutman GA, Navon R (1999) hKCa3/KCNN3 potassium channel gene: Association of longer CAG repeats with schizophrenia in Israeli Ashkenazi Jews, expression in human tissues and localization to chromosome 1q21. Mol Psychiatry 4:254–260

Ellegren H, Moore S, Robinson N, Byrne K, Ward W, Sheldon BC (1997) Microsatellite evolution—A reciprocal study of repeat lengths at homologous loci in cattle and sheep. Mol Biol Evol 14:854–860

Ellegren H, Primmer CR, Sheldon BC (1995) Microsatellite 'evolution': directionality or bias? Nat Genet 11:360–362

Enard W, Przeworski M, Fisher SE, Lai CS, Wiebe V, Kitano T, Monaco AP, Pääbo S (2002) Molecular evolution of FOXP2, a gene involved in speech and language. Nature 418:869–872

Figueroa KP, Chan P, Schols L, Tanner C, Riess O, Perlman SL, Geschwind DH, Pulst SM (2001) Association of moderate polyglutamine tract expansions in the slow calcium-activated potassium channel type 3 with ataxia. Arch Neurol 58:1649–1653

Fu YH, Kuhl DP, Pizzuti A, Pieretti M, Sutcliffe JS, Richards S, Verkerk AJ, Holden JJ, Fenwick RG Jr., Warren ST, Oostra BA, Nelson DL, Caskey CT (1991) Variation of the CGG repeat at the fragile X site results in genetic instability: resolution of the Sherman paradox. Cell 67:1047–1058

Gagneux P, Varki A (2001) Genetic differences between humans and great apes. Mol Phylogenet Evol 18:2–13

Garza JC, Slatkin M, Freimer NB (1995) Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size. Mol Biol Evol 12:594–603

González-Cabo P, Sanchez MI, Canizares J, Blanca JM, Martínez-Arias R, DeCastro M, Bertranpetit J, Palau F, Molto MD, deFrutos R (1999) Incipient GAA repeats in the primate Friedreich ataxia homologous genes. Mol Biol Evol 16:880–883

Hacia JG (2001) Genome of the apes. Trends Genet 17:637–645

Hedrick PW, Thomson G (1983) Evidence for balancing selection at HLA. Genetics 104:449–456

Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335:167–170

Ingman M, Kaessmann H, Pääbo S, Gyllensten U (2000) Mitochondrial genome variation and the origin of modern humans. Nature 408:708–713

Jodice C, Giovannone B, Calabresi V, Bellocchi M, Terrenato L, Novelletto A (1997a) Population variation analysis at nine loci containing expressed trinucleotide repeats. Ann Hum Genet 61(Pt 5):425–438

Jorde LB, Watkins WS, Bamshad MJ, Dixon ME, Ricker CE, Seielstad MT, Batzer MA (2000) The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. Am J Hum Genet 66:979–988

Justice CM, Den Z, Nguyen SV, Stoneking M, Deininger PL, Batzer MA, Keats BJ (2001) Phylogenetic analysis of the Friedreich ataxia GAA trinucleotide repeat. J Mol Evol 52:232–238

Kaessmann H, Wiebe V, Pääbo S (1999) Extensive nuclear DNA sequence diversity among chimpanzees. Science 286:1159–1162

Kaessmann H, Wiebe V, Weiss G, Pääbo S (2001) Great ape DNA sequences reveal a reduced diversity and an expansion in humans. Nat Genet 27:155–156

Kawaguchi Y, Okamoto T, Taniwaki M, Aizawa M, Inoue M, Katayama S, Kawakami H, Nakamura S, Nishimura M, Akiguchi I (1994) CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1. Nat Genet 8: 221–228

Koob MD, Moseley ML, Schut LJ, Benzow KA, Bird TD, Day JW, Ranum LP (1999) An untranslated CTG expansion causes a novel form of spinocerebellar ataxia (SCA8). Nat Genet 21: 379–384

Lai CS, Fisher SE, Hurst JA, Vargha-Khadem F, Monaco AP (2001) A forkhead-domain gene is mutated in a severe speech and language disorder. Nature 413:519–523

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

Messier W, Stewart CB (1997) Episodic adaptive evolution of primate lysozymes. Nature 385:151–154

Nagafuchi S, Yanagisawa H, Sato K, Shirayama T, Ohsaki E, Bundo M, Takeda T, Tadokoro K, Kondo I, Murayama N (1994) Dentatorubral and pallidoluysian atrophy expansion of an unstable CAG trinucleotide on chromosome 12p. Nat Genet 6: 14–18

Noda R, Kim CG, Takenaka O, Ferrell RE, Tanoue T, Hayasaka I, Ueda S, Ishida T, Saitou N (2001) Mitochondrial 16S rRNA sequence diversity of hominoids. J Hered 92:490–496

Nolin SL, Brown WT, Glicksman A, Houck GE Jr., Gargano AD, Sullivan A, Biancalana V, Brondum-Nielsen K, Hjalgrim H, Holinski-Feder E, Kooy F, Longshore J, Macpherson J, Mandel JL, Matthijs G, Rousseau F, Steinbach P, Vaisanen ML, Von Koskull H, Sherman SL (2003) Expansion of the fragile X CGG repeat in females with premutation or intermediate alleles. Am J Hum Genet 72:454–464

Orr HT, Chung MY, Banfi S, Kwiatkowski TJ, Jr., Servadio A, Beaudet AL, McCall AE, Duvick LA, Ranum LP, Zoghbi HY (1993) Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. Nat Genet 4: 221–226

Pulst SM, Nechiporuk A, Nechiporuk T, Gispert S, Chen XN, Lopes-Cendes I, Pearlman S, Starkman S, Orozco-Diaz G, Lunkes A, DeJong P, Rouleau GA, Auburger G, Korenberg JR, Figueroa C, Sahba S (1996) Moderate expansion of a normally biallelic trinucleotide repeat in spinocerebellar ataxia type 2. Nat Genet 14: 269–276

Rubinsztein DC, Amos W, Leggo J, Goodburn S, Jain S, Li SH, Margolis RL, Ross CA, Ferguson-Smith MA (1995a) Microsatellite evolution—Evidence for directionality and variation in rate between species. Nat Genet 10:337–343

Rubinsztein DC, Leggo J, Coetzee GA, Irvine RA, Buckley M, Ferguson-Smith MA (1995b) 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

Ruvolo M (1997) Molecular phylogeny of the hominoids: Inferences from multiple independent DNA sequence data sets. Mol Biol Evol 14:248–265

Saleem Q, Anand A, Jain S, Brahmachari SK (2001) The polyglutamine motif is highly conserved at the Clock locus in various organisms and is not polymorphic in humans. Hum Genet 109:136–142

Schneider S, Roessli D, Excoffier L (2000) A software for population genetics data analysis Genetics and Biometry Laboratory, University of Geneva, Switzerland

Sokal RR, Rohlf FJ (1995) Biometry WH Freeman, New York,

Watkins WS, Bamshad M, Jorde LB (1995) Population genetics of trinucleotide repeat polymorphisms. Hum Mol Genet 4:1485–1491

Webster MT, Smith NG, Ellegren H (2002) Microsatellite evolution inferred from human-chimpanzee genomic sequence alignments. Proc Natl Acad Sci USA 99:8748–8753

Wise CA, Sraml M, Rubinsztein DC, Easteal S (1997) Comparative nuclear and mitochondrial genome diversity in humans and chimpanzees. Mol Biol Evol 14:707–716

Yue S, Serra HG, Zoghbi HY, Orr HT (2001) The spinocerebellar ataxia type 1 protein, ataxin-1, has RNA-binding activity that is inversely affected by the length of its polyglutamine tract. Hum Mol Genet 10:25–30

Zhuchenko O, Bailey J, Bonnen P, Ashizawa T, Stockton DW, Amos C, Dobyns WB, Subramony SH, Zoghbi HY, Lee CC (1997) Autosomal dominant cerebellar ataxia (SCA6) associated with small polyglutamine expansions in the alpha 1A-voltage-dependent calcium channel. Nat Genet 15: 62–69