European Journal of Human Genetics (2006) 14, 932–940

www.nature.com/ejhg

ARTICLE

A multistep mutation mechanism drives the evolution of the CAG repeat at *MJD/SCA3* locus

Sandra Martins*[,1,2], Francesc Calafell[3], Virginia CN Wong[4], Jorge Sequeiros[5] and António Amorim[1,2]

[1]*IPATIMUP – Instituto de Patologia e Imunologia Molecular da Universidade do Porto, Porto, Portugal;* [2]*Faculdade de Ciências, Universidade do Porto, Porto, Portugal;* [3]*Unitat de Biologia Evolutiva, Facultat de Ciències de la Salut i de la Vida, Universitat Pompeu Fabra, Barcelona, Spain;* [4]*Department of Paediatrics and Adolescent Medicine, The University of Hong Kong, Hong Kong, China;* [5]*UnIGENe – IBMC and ICBAS, Universidade do Porto, Porto, Portugal*

Despite the intense debate around the repeat instability reported on the large group of neurological disorders caused by trinucleotide repeat expansions, little is known about the mutation process underlying alleles in the normal range that, ultimately, expand to pathological size. In this study, we assessed the mutation mechanisms by which wild-type Machado–Joseph disease (MJD) alleles have been generated throughout human evolution. Haplotypes including the CAG repeat, six intragenic SNPs and four flanking microsatellites were analysed in 431 normal chromosomes of European, Asian and African origin. A bimodal CAG repeat length frequency distribution was found in the four most frequent wild-type lineages (H1-GCGGCA; H2-GTGGCA; H3-TTAGAC and H4-TTACAC). Based on flanking microsatellite haplotypes, the variance calculated by analysis of molecular variance between modal $(CAG)_n$ alleles was little or null in lineages H1, H2 and H4, as were the pairwise differences. Moreover, genetic distances among all the alleles from each lineage did not reflect the allele sizes differences, as expected if a stepwise mutation model was the main process of evolution. On the contrary, when exposed in maximum parsimonious phylogenetic trees, a large number of mutation steps separated same-size alleles, whereas several microsatellite haplotypes were shared by modal CAGs. In conclusion, our results suggest that the main mutation mechanism occurring in the evolution of the polymorphic CAG region at *MJD/SCA3* locus is a multistep one, either by gene conversion or DNA slippage; repeats with 14, 21, 23 and 27 CAGs are the main alleles involved in this process.
European Journal of Human Genetics (2006) **14**, 932–940. doi:10.1038/sj.ejhg.5201643; published online 17 May 2006

Keywords: Machado–Joseph disease; multistep mutation model; gene conversion; haplotype

Introduction

The discovery of human genetic disorders caused by repeat expansions has brought a new interest regarding the evolutionary processes that underlie these repeat tracts.[1]

Machado–Joseph disease (MJD), also called spinocerebellar ataxia type 3 (SCA3), is a late-onset neurodegenerative disorder with dominant inheritance, caused by a CAG repeat expansion located in exon 10 of the *ATXN3* gene. When expanded alleles were first described, the found range was 68–79 CAGs, but these limits have continuously been broadened with new results reporting the high intergenerational instability observed mainly during paternal transmissions.[2–5] The molecular mechanism

*Correspondence: S Martins, IPATIMUP – Instituto de Patologia e Imunologia Molecular da Universidade do Porto, Rua Dr. Roberto Frias s/n, Porto 4200-465, Portugal.
Tel: +351 22 5570700; Fax: +351 22 5570799;
E-mail: smartins@ipatimup.pt
Received 23 November 2005; revised 7 February 2006; accepted 28 March 2006; published online 17 May 2006

often proposed as causative of the triplet repeat instability is the aberrant replication of DNA owing to the formation of unusual conformations like slipped structures, triplexes, flexible and writhed DNA.[6–8] Other studies have shown homologous recombination inducing CAG/CTG repeat contractions and expansions during mitotic gene conversion in yeast and in bacteria.[9,10]

Contrarily to the intensive investigation on expanded alleles, little research has been aimed at mutation processes underlying the evolution of normal alleles. Therefore, the highly polymorphic content of normal MJD alleles, usually varying between 12 and 44 repeats, has been assumed as explainable by the simple commonly observed process of microsatellite evolution – the stepwise mutation model (SMM). Nevertheless, when considering the gap between normal and expanded ranges, a strict SMM, in which alleles mutate by the gain or loss of one repeat unit (under a slippage mutation rate of $\sim 10^{-4}$/gamete/generation), seems insufficient to explain the evolution of this polyglutamine tract. Among a variety of species, such as *Pan troglodytes*, *Gorilla gorilla*, *Mus musculus* and *Gallus gallus*, the repetitive region in the *ATXN3* gene is conserved, although much shorter than their human homologues. By comparing anthropoid primate species, it was suggested that the expansion has occurred in the human lineage alone, after diverging from all other hominoid lineages.[11] Moreover, because unrelated genes at different locations have expanded in the higher primates, it was postulated that expansion events could be affected by the product of other genes, possibly the same that control the rates of expansion of microsatellites and simple sequences. More recently, a relationship between variability levels of normal repeats and expansion potential was suggested, as larger variances in humans, compared to chimpanzees, gorillas and orangutans, were only observed for the expanding loci.[12]

In the present study, to gain insight into the dynamics of the MJD trinucleotide repeat in human evolution, we have identified the wild-type lineages in three continents, and have placed them in a haplotype background of stable (SNP) and fast-evolving (STR) polymorphisms. The use of a stable SNP haplotype background allows disentangling identity-by-descent from identity-by-state at the CAG repeat, which is essential if the mutation relationships among CAG repeat alleles are to be determined. Under an SMM, it is expected that, on average, alleles differing by a larger number of repeat units have been diverging for a longer time, and, therefore, their haplotype backgrounds have been diverging as well. We computed several measures of divergence around MJD alleles, and found a complex result that is hard to reconcile with a strict SMM, which points to multistep events occurring frequently between some CAG alleles, or alternatively high rates of recombination and/or gene conversion.

Subjects and methods

DNA samples

Peripheral blood samples of unrelated individuals and some parent–child pairs were collected, after their informed consent, in order to extract genomic DNA by standard procedures. A total of 304 individuals were analysed, including 216 of European (Portuguese), 50 of Asian (Chinese) and 38 of African (Mozambican and Angolan) origin. As a result, 291 phase-known chromosomes (265 of European and 26 of African origin) and 140 unrelated chromosomes (12 of European, 100 of Asian and 28 of African population) were analysed. Two *P. troglodytes* and one *G. gorilla* were genotyped for the SNPs in which ancestral alleles were not known.

Polymorphic markers

The biallelic markers studied, GTT^{527}/GTC^{527}, A^{669}TG/G^{669}TG, C^{987}GG/G^{987}GG, TAA^{1118}/TAC^{1118} and C^{1178}/A^{1178}, were previously described.[13–15] The IVS6-30G>T was newly discovered by us, while typing the GTT^{527}/GTC^{527} SNP. The four flanking microsatellites were selected based on their consensus size (2–5), copy number (>8) and percent matches (>80%) of the putative polymorphic repeats found by the Tandem Repeats Finder software http://tandem.bu.edu/trf/trf.html.[16] The search was carried out in the genomic contig that includes the *ATXN3* gene (AL049872) and in the flanking AL590328, AL121773 and AL133240 contigs, according to the Ensembl Human Genome Browser (http://www.ensembl.org/).

Genotyping

An ARMS-PCR, encompassing the CAG repeat and the biallelic markers C^{987}GG/G^{987}GG and TAA^{1118}/TAC^{1118}, was performed, using the primers MJD52 and either ASP3′ or ASP4′, labelled with the fluorescent tags 6-FAM and TET, respectively. Forward primers differ from the described ASP3 and ASP4[17] at the second 3′ base, where a C>G mismatch was introduced to increase the allele-specificity. The amplification reaction was carried out in a total volume of 12.5 μl, with 0.4 μM of each primer, 200 μM dNTPs, 1 mM $MgCl_2$, 20 mM ammonium sulphate, 1.84% DMSO and 1 U of *Taq* polymerase. Hot-start amplification conditions were as follows: initial denaturation step at 94°C, for 7 min; 30 cycles consisting of 94°C, for 1 min, annealing condition, and 72°C, for 1 min; and a final extension of 7 min, at 72°C (details regarding amplification and genotyping supplied as Supplementary material – S1).

DNA fragment length analysis was performed on the ABI-Prism 310 Genetic Analyser laser-induced fluorescence capillary electrophoresis system (Applied Biosystems, Foster City, CA, USA), using the GeneScan Analysis 3.1 software with TAMRA-500 as standard. Simultaneously, the TAA^{1118}/TAC^{1118} SNP was genotyped based on the fluorescence emitted from the allelic-specific primers. A subsequent restriction endonuclease digestion was carried out

with the enzyme *Msp*A1 I, to assess the allelic phase of SNP $\underline{C}^{987}GG/\underline{G}^{987}GG$ with the two previously typed polymorphisms.

The $GT\underline{T}^{527}/GT\underline{C}^{527}$, $\underline{A}^{669}TG/\underline{G}^{669}TG$ and C^{1178}/A^{1178} SNPs were genotyped by PCR-SSCP. Amplification conditions were as described previously, and the denatured PCR products were submitted to electrophoresis, at 4°C, in a $T_{12}C_5$ polyacrylamide gel. Products were visualized by silver staining.

Several different allele sizes of each polymorphic locus were sequenced with the Big Dye Terminator Cycle Kit (Applied Biosystems, Foster City, CA, USA), after purification with Microspin S-300 HR columns (Pharmacia), in order to correlate allele size estimated by GeneScan analysis and the exact number of repeats of studied microsatellites and CAG repeat. The post-reaction purification was performed in SigmaSpin Clean-Up columns (Sigma) and products run in an ABI 3100 sequencer (Applied Biosystems, Foster City, CA, USA). The results were analysed using the Data Collection software.

Data analysis

Allele frequencies were estimated by direct counting. The PHASE 2.0 software (www.stat.washington.edu/stephens/software.html) was used to infer haplotypes from genotypic data, whenever the complete allelic phase of the 11 analysed loci was not directly inferred. Three separate analyses were performed, one per population, including in each case the phase-known haplotypes obtained from 291 analysed family chromosomes and allele-specific amplifications.

Phylogenetic networks of SNP-based haplotypes, as well as of microsatellite variation for the four most frequent lineages were performed using the Network 4.0.1.6 software (http://www.fluxus-technology.com). A reduced median followed by a median-joining network was calculated to resolve some of the reticulation at microsatellite loci. In all calculations, ε was set to zero, but different weights were assigned to microsatellites, according to their variance in allele repeat size, in an inverse proportional ratio. After identifying the wild-type MJD lineages, the analysis of molecular variance (AMOVA), performed with the Arlequin 2.000 software (http://anthro.unige.ch/arlequin), allowed the comparison of molecular variation between and within the modal CAG alleles for the three most frequent lineages.[18]

To assess the molecular vicinity among the different CAG alleles of each lineage, a pairwise analysis of the flanking haplotypes was also performed with Arlequin 2.000 software, applying the sum of squared size difference (R_{ST}) as the distance method. This way, the estimation of evolutionary distance between each pair of haplotypes was calculated taking into account the number of presumed single-step mutation steps between the corresponding STR alleles.

Results

Wild-type lineages

Genotyping of six intragenic SNPs (Table 1) in the three populations showed 11 MJD lineages in non-expanded chromosomes, although seven of them were extremely rare (Figure 1). The haplotype H1 (GCGGCA) was the most frequent among European (51.3%) and African (40.7%)

Table 1 Description of the polymorphic markers used in the haplotype study and their distance from the CAG repeat at the *MJD/SCA3* locus

Polymorphism	*Location*	*Distance from the $(CAG)_n$ (bp)*
$(TAT)_n$	AL133240	223 315
$(CA)_n$	AL133240	190 946
IVS6-30G>T (ss35527515)	Intron 6	12 447
$GT\underline{T}^{527}/GT\underline{C}^{527}$ (ss24128207)	Exon 7	12 201
$\underline{A}^{669}TG/\underline{G}^{669}TG$ (ss14386151)	Exon 8	11 400
$(CAG)_n$	Exon 10	—
$\underline{C}^{987}GG/\underline{G}^{987}GG$ (ss21217642)	Exon 10	1
$TA\underline{A}^{1118}/TA\underline{C}^{1118}$ (ss10756371)	Exon 10	131
C^{1178}/A^{1178} (ss10758189)	Intron 10	191
$(AC)_n$	AL049872	20 817
$(GT)_n$	AL590328	189 857

NCBI assay IDs for SNPs are in brackets.
Distance from the CAG repeat is according to the complete sequence of Chromosome 14 at the NCBI (NT_026437).

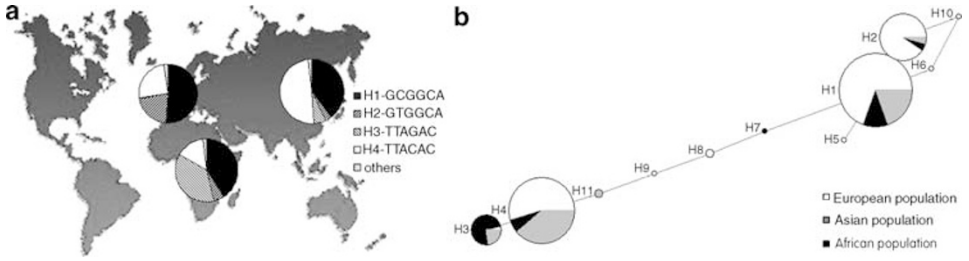

Figure 1 SNP-based haplotypes at *MJD/SCA3* locus. (**a**) Haplotype frequencies in European, Asian and African populations. (**b**) Median-joining network of the wild-type MJD lineages, where the circle area is proportional to frequency and branch length proportional to the number of mutations.

populations; in Asia, although H1 reached 40%, the haplotype H4 (TTACAC) was the most prevalent (49%); on the other hand, H4 was present in 24.9 and 14.8% of the European and African chromosomes, respectively. This is another example of a *yin-yang* pattern in the two most frequent haplotypes.[19] The other two most frequent lineages H2 (GTGGCA) and H3 (TTAGAC) derived from the H1 and H4 haplotypes by just one mutation step at SNPs GT$\underline{T}^{527}$/GT$\underline{C}^{527}$ and $\underline{C}^{987}$GG/$\underline{G}^{987}$GG, respectively (Figure 1b). Concerning their geographical distribution, H3 was the second most frequent haplotype in Africa (37%), but was shared only by 6% of the Asian and 0.4% of the European chromosomes, whereas H2 reached 20.9% in the European, 5.5% in African and 3% in Asian populations. The ancestral haplotype TTGGCC was never found among the human chromosomes analysed.

The CAG frequencies showed always a bimodal distribution inside each lineage, with distinct modal values for each one (Figure 2). Within H1 lineage, 82.4% of the chromosomes carried either 14 or 23 CAGs, whereas in the H2 haplotype, the vast majority harboured alleles with 23 or 27 repeats. A more scattered distribution was found in H3 and H4 lineages, although a gap of five CAGs (with only 4.8% of the chromosomes) was still observed between H4 modal alleles. It was also in this lineage that the largest normal alleles were observed, reaching 10.3% for alleles larger than 30 CAGs.

The bimodality of the lineages H1, H2 and H4, regarding the CAG distribution according to the population of origin, was only observed in the European chromosomes. Moreover, in a worldwide distribution, alleles of Asian origin had only modal values at the H1 and H4 lineages, carrying 14 and 27 CAGs, respectively.

Molecular distances among CAG alleles

To study in more detail the evolution of CAG alleles within each lineage, flanking STR-based haplotypes were analysed in the three populations. Repetitive motifs and distances from the CAG repeat are described in Table 1. The microsatellite that is most distant from the CAG repeat is 223 kb away; this translates to a genetic distance of $\sim$0.3 cM, taking into account the recombination rate of 1.41 cM/Mb estimated for the nearest marker D14S1015.[20] Results from sequencing have shown three pure and one complex interrupted polymorphic marker (allelic frequencies for each population are reported in Supplementary material – S2). The complex $(GT)_n$ repeat presented perfect $(GT)_{14}$, $(GT)_{15}$, $(GT)_{17}$ and $(GT)_{18}$ alleles, whereas

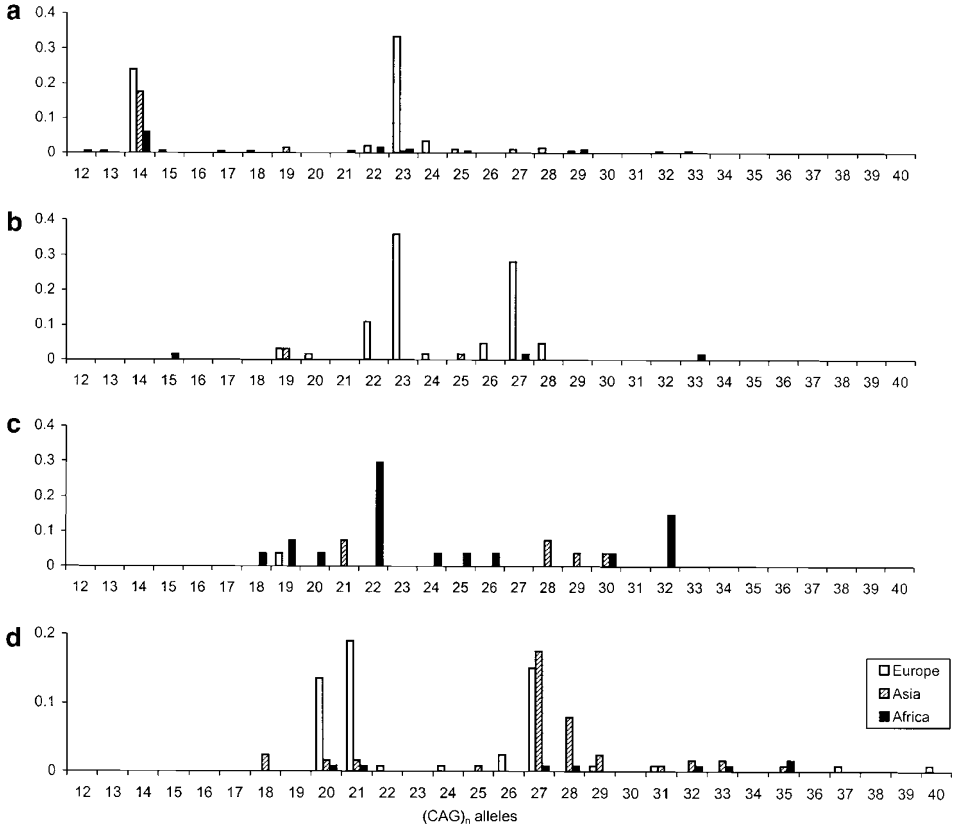

Figure 2 $(CAG)_n$ allele frequencies at *MJD/SCA3* locus in European, Asian and African control populations, observed in (**a**) H1- GCGGCA, (**b**) H2- GTGGCA, (**c**) H3- TTAGAC and (**d**) H4- TTACAC lineages.

Table 2 AMOVA comparing the percentage of variation between the two modal $(CAG)_n$ alleles for the most frequent wild-type lineages at *MJD/SCA3* locus

	Variation between modal alleles (%)		
	H1- GCGGCA	*H2- GTGGCA*	*H4- TTACAC*
All haplotypes	0 ($n=168$)	0 ($n=42$)	6.17 ($n=69$)
Phase-known haplotypes	0 ($n=65$)	0 ($n=20$)	20.57 ($n=36$)
STR $(AC)_n$	5.51 ($n=168$)	0 ($n=42$)	27.24 ($n=69$)

'n' denotes the number of modal alleles analysed in each case.

interrupted alleles 13, 19 and 16 had the motifs $(GT)_{4;10}$ AT $(GT)_5$ GG $(GT)_2$ and $(GT)_{13}$ GG $(GT)_2$, respectively.

The diversity at the linked STRs was used to assess the divergence between the modal CAG repeats at each lineage. Based (1) on all haplotypes obtained by segregation analysis and PHASE reconstruction, (2) on phase-known haplotypes and (3) on the closest microsatellite analysed, the amount of divergence was quantified by means of AMOVA for the three most frequent MJD lineages (Table 2). This way, we removed the possible source of bias owing to incorrect haplotype reconstructions by PHASE, or recombination. With the exception of a slight variation in H4, similar proportions of variation between modal alleles were obtained from the three analyses. Actually, for lineage H2, the fraction of genetic variation of the flanking STRs, explained by differences between modal alleles, was always 0; the same was true for modal alleles of H1, when all or only phase-known haplotypes were considered, and 5.51 when the AMOVA was performed with only the $(AC)_n$ microsatellite.

The stepwise model was also tested comparing the observed molecular distances among CAG alleles and the predicted distributions under this model. In fact, the distance between the STR haplotypes linked with each CAG allele is expected to increase with the difference in the number of repeats to the modal alleles. However, when comparing pairwise differences among CAG alleles, R_{ST} values did not increase proportionally with the repeat number distance to a modal allele (Figure 3). In lineage H1, for example, the molecular distance between the modal allele 14 and the neighbouring 13 and 15 alleles was 14.2 and 29.6, respectively, but 0.0 when compared to the other modal allele (with 23 CAGs). Still within this lineage, allele 28 shared also similar flanking haplotypes with both modal alleles ($R_{ST}=0.0$). In the same way, modal alleles 23 and 27 of lineage H2 were molecularly more similar to each other ($R_{ST}=0.0$) than to other alleles just one or two CAG units apart. In lineage H4, although not as evident as in H1 or H2, we have still observed low R_{ST} values between modal alleles, whereas the opposite happened in lineage H3, in which flanking haplotypes were shared, not by modal repeats, but by allele 32 and alleles with 28 and 30 CAGs. Analysing only phase-known haplotypes (inferred by segregation analysis), similar genetic distances were obtained, corroborating the results (data on Supplementary material – S3).

Accordingly, in phylogenetic trees constructed for each SNP-based lineage, representing the molecular distances among the STR-based haplotypes, no subbranches arose according to the CAG allele size (Figure 4). A homogeneous distribution of the modal CAG alleles throughout the different clades of the tree, as well as the sharing of haplotypes between them, was observed in all lineages. Moreover, alleles harbouring the same number of CAG repeats were separated by a huge number of mutation steps, which suggests that the molecular distance between them is much higher than between this allele and all the others in between.

Discussion

To address the question of CAG repeat instability, we dissected the evolution of wild-type MJD alleles, showing that the expectations of a single-step mutation model were not fulfilled. Conversely, our results are compatible with the hypothesis that the frequency distribution of CAG alleles has been shaped by a multistep mutation mechanism.

Analysing samples from three major human population groups (European, Asian and African), MJD alleles within the normal range followed a trimodal distribution, as in other worldwide series, displaying almost no gaps;[21,22] however, when distributions are analysed separately by population of origin, different patterns emerge. Chinese individuals carried either the alleles 14 or 27 in 58% of the chromosomes, whereas alleles with 23 CAGs added up only to 1%, as it was found in Japanese individuals.[23] On the other hand, the most scattered distribution was observed in Africans, also in agreement with that described previously.[24,25] This heterogeneity among populations has been poorly studied, as it was considered exclusively the consequence of demographic events. When analysing the alleles by SNP-based haplotype lineage, however, we have found not a modal CAG size in a normal distribution (as expected, if diversity had been created by stepwise mutations from the ancestral CAG allele where the new lineage had newly arisen), but instead a bimodal distribution with

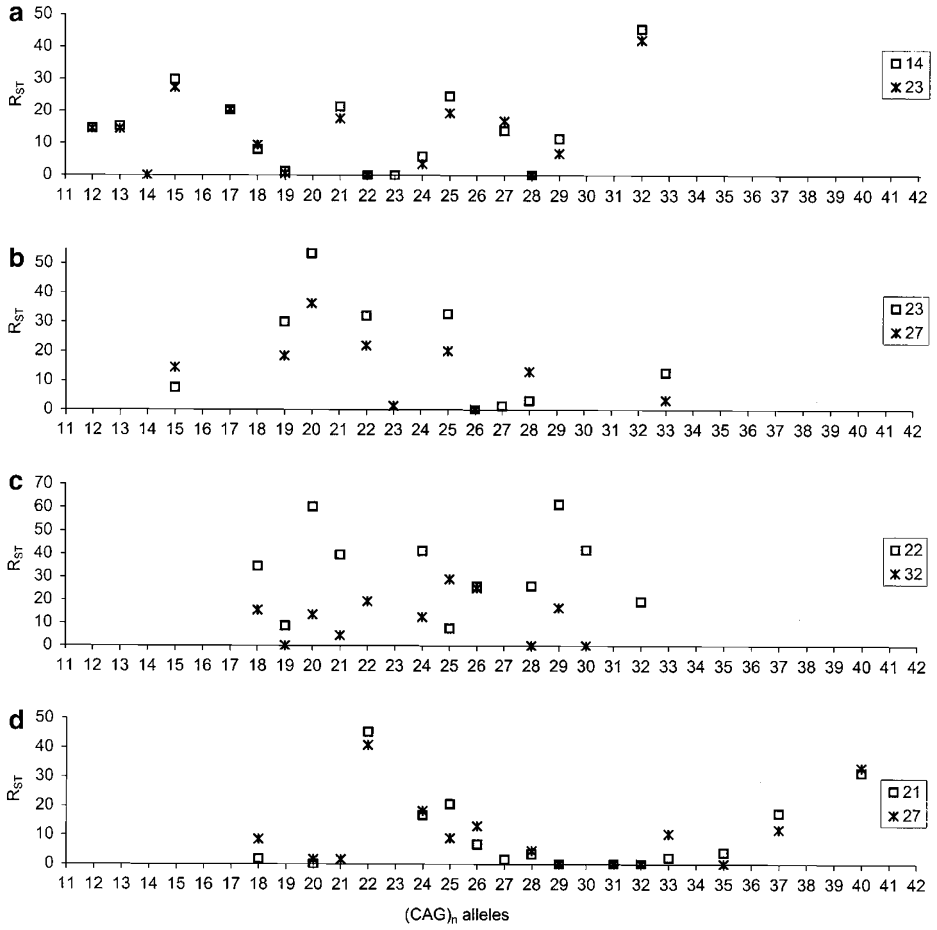

Figure 3 Genetic distance between the two CAG modal alleles and the rest of the alleles found within the same wild-type MJD lineage. (**a**) H1-GCGGCA, (**b**) H2- GTGGCA, (**c**) H3- TTAGAC and (**d**) H4- TTACAC. Modal alleles and respective symbols for associated R_{ST} distances are shown in the box, at the upper right corner at the respective graphic.

modal alleles separated by a gap of three to nine repeat units. The sum of inter-modal alleles frequencies was less than 7.8% in the SNP-defined lineages H1, H2 and H4. This previous definition is crucial in CAG allele comparisons, as completely different evolutionary fates can be hidden under the same CAG allele size. Therefore, while associating frequencies of large normal alleles to disease prevalence rates, as carried out previously,[21,23] SNP-based haplotypes should not be ignored as only alleles (normal or expanded) of the same lineage are comparable. Moreover, large normal alleles in one lineage may be just average in another one, as different ranges are often found, as we show in this study. For the same reason, the use of CAG allele's class intervals can be misleading, and haplotype backgrounds become a fundamental tool to investigate the actual origin of expansions from the normal range.

Hence, results from the distribution of wild-type MJD alleles, according to their lineage, provided us the first evidence that a multistep mutation mechanism must have occurred at the *MJD/SCA3* locus. The fact that very distant alleles, in terms of length, have similar frequencies in the same lineage suggests that these events were not sporadic, but a major process in the locus evolution. To test the hypothesis of replicative-like events, such as duplications or constitutive tandem expansions, underlying this locus diversity, we counted the number of CAGs between modal values, but no identical or multiplicative numbers were observed, even eliminating the first six interrupted repeats. Therefore, the intriguing properties of this distribution encouraged us to unravel the molecular backgrounds of modal alleles and their neighbouring size alleles. If the mutations that have separated the modal alleles were single-step, the time elapsed would imply that the haplotype backgrounds for each modal CAG repeat allele have diverged noticeably. AMOVA results, pairwise comparisons and phylogenetic analyses not only failed to correlate the genetic distances with the allele size difference but also showed insignificant genetic distances between modal alleles, as well as a large number of

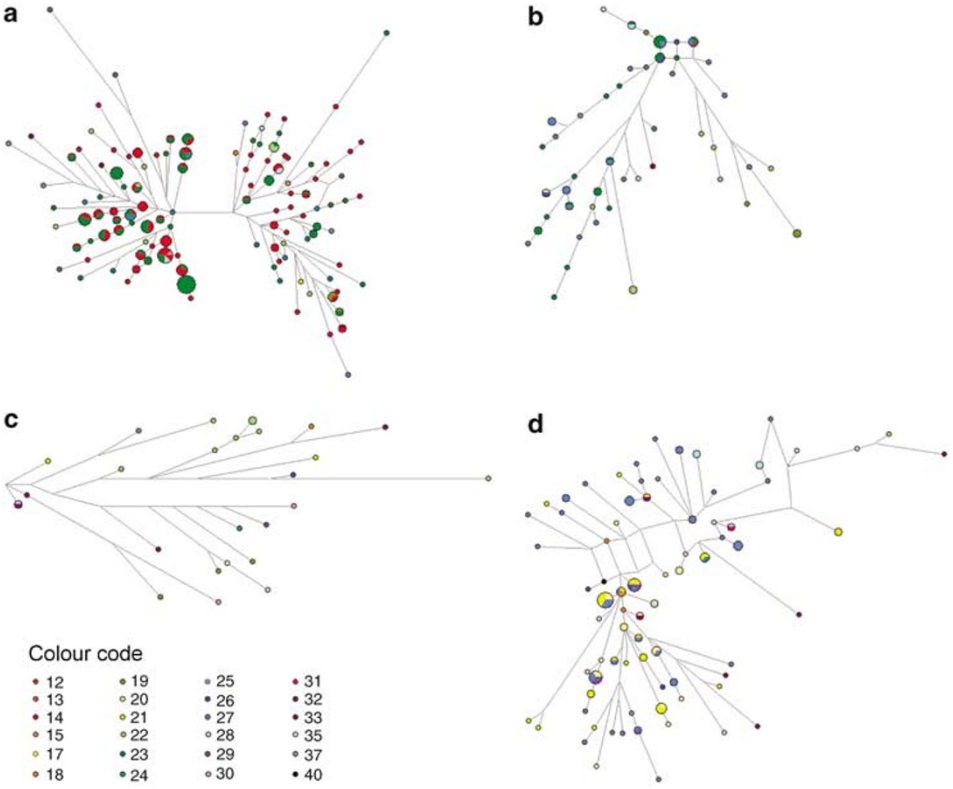

Figure 4 Median-joining network of the STR-based haplotypes for (**a**) H1- GCGGCA, (**b**) H2- GTGGCA, (**c**) H3- TTAGAC and (**d**) H4- TTACAC wild-type MJD lineages. Circle area is proportional to frequency and branch length is proportional to the number of mutations. Allele sizes are differently coloured and colour code numbers refer to the number of repeats.

mutation steps separating alleles with the same size. Anyway, we need to exclude a possible spurious explanation for these results: as some of the haplotypes were estimated by statistical inference methods, it could be that the algorithm wrongfully assorted STR alleles with respect to the other haplotype markers. However, when just family-resolved haplotypes were used, examples of modal haplotypes sharing the same STR background were found, and similar results were obtained by both AMOVA and pairwise analysis. Another explanation could consist in massive double recombination events, occurring in H1, H2 and H4 lineages, between modal alleles. However, no such events were found in the analysed nuclear families. Thus, multistep events, either by DNA slippage or gene conversion, seem a more plausible hypothesis for the major mutation mechanism occurring between modal alleles, whereas the rest of the alleles may have derived by stepwise mutations. Taking into account that recombined SNP-based haplotypes were rarely found, hot-spot sites for these gene conversion events could either be (1) flanking the 12.6 kb region that encompasses the CAG repeat and all the intragenic SNPs analysed or (2) somewhere between the $\underline{A}^{669}TG/\underline{G}^{669}TG$ SNP and the end of the CAG array. It is noteworthy that the bimodality, suggested as the better evidence of the multistep mutation events, occurs mainly among European chromosomes, in lineages H1, H2 and H4. Considering that the variation of STR-defined haplotypes among CAG alleles was always observed within a previously identified lineage, however, these population differences cannot be explained by any demographic effects. Further studies on putative *cis*- and *trans*-acting modifier factors will be needed to gain insight into this matter.[26]

Recently, gene conversion was proposed as the mechanism involved in the origin of a rare intermediate allele of MJD, as it had a flanking SNP-based haplotype (A–C–A) commonly observed in large alleles and, simultaneously, a tract variant at the sixth repeat (CAA instead of CAG), only observed in smaller alleles that were significantly associated with a different SNP-based haplotype (G–G–C).[27] As for other CTG·CAG tracts, gene conversion was reported in a few clinical cases of myotonic dystrophy (MD)[28,29] and Huntington disease,[30] whereas other alternative models for the evolution of wild-type alleles have been hypothesized. In Friedreich ataxia, two duplication events were suggested to explain the distribution of GAA repeat sizes observed in normal chromosomes.[31] The authors proposed that the first event occurred presumably only once in Africa, from a

small normal allele containing eight or nine GAAs to a large normal allele with 16 or 18 triplets. Then, after single repeat insertions/deletions, derived from SMM, one or more chromosomes with 12 to 25 GAAs would have migrated to Europe, the Middle East or both, before the second duplication event has given rise to the large normal alleles with more than 30 GAAs. In MD, a multistep mechanism was also proposed, under which a very rare (possibly also a single one) transition had occurred from a $(CTG)_5$ allele to an allele with 19 to 30 repeats, whereas the heterogeneous class of $(CTG)_{19-30}$ alleles was suggested as a reservoir for recurrent MD mutations.[32] On the other hand, in the case of SCA1 and SCA2, studying the distribution of interruptions within the tracts, it was suggested that the dynamic mutation of *ATXN1* and *ATXN2* genes initiated from the expansion of long pure repeat tracts without the prior loss of interruptions.[33]

In conclusion, our results strongly support a multistep mutation model underlying the evolution of the CAG alleles at the *MJD/SCA3* locus. It would be interesting, using the same approach, to extend this study to MJD families in order to confirm if the same type of mutation mechanism is also underlying the expanded alleles and *de novo* mutational events. If so, further analysis of modal alleles identified here could give us a clue on nonfamiliar cases of the disease. Additionally, studying other expanding repeat loci, one could clarify whether this finding is restricted to this locus or a landmark of other polyglutamine disease expansions.

Acknowledgements

We thank Dr Albertino Damasceno and Dr Benilde Soares of the Eduardo Mondlane University (Maputo) for kindly providing the Mozambican samples. This work was partially supported by FCT (Fundação para a Ciência e Tecnologia), through research grant POCTI (Programa Operacional Ciência, Tecnologia e Inovação) and the scholarship SFRH/BD/8880/2002 attributed to S Martins.

References

1 Andrés AM, Lao O, Soldevila M, Calafell F, Bertranpetit J: Dynamics of CAG repeat loci revealed by the analysis of their variability. *Hum Mutat* 2003; **21**: 61–70.
2 Kawaguchi Y, Okamoto T, Taniwaki M *et al*: CAG expansions in a novel gene for Machado–Joseph disease at chromosome 14q32.1. *Nat Genet* 1994; **8**: 221–228.
3 Takiyama Y, Igarashi S, Rogaeva E *et al*: Evidence for intergenerational instability in the CAG repeat in the MJD1 gene and for conserved haplotypes at flanking markers amongst Japanese and Caucasian subjects with Machado–Joseph disease. *Hum Mol Genet* 1995; **4**: 1137–1146.
4 Igarashi S, Takiyama Y, Cancel G *et al*: Intergenerational instability of the CAG repeat of the gene for Machado–Joseph disease (MJD1) is affected by the genotype of the normal chromosome: implications for the molecular mechanisms of the instability of the CAG repeat. *Hum Mol Genet* 1996; **5**: 923–932.
5 Maciel P, Gaspar C, Guimarães L *et al*: Study of three intragenic polymorphisms in the Machado–Joseph disease gene (MJD1) in relation to genetic instability of the $(CAG)_n$ tract. *Eur J Hum Genet* 1999; **7**: 147–156.
6 Usdin K, Woodford KJ: CGG repeats associated with DNA instability and chromosome fragility form structures that block DNA synthesis *in vitro*. *Nucleic Acids Res* 1995; **23**: 4202–4209.
7 Pearson CE, Sinden RR: Alternative structures in duplex DNA formed within the trinucleotide repeats of the myotonic dystrophy and fragile X loci. *Biochemistry* 1996; **35**: 5041–5053.
8 Wells RD, Parniewski P, Pluciennik A, Bacolla A, Gellibolian R, Jaworski A: Small slipped register genetic instabilities in *Escherichia coli* in triplet repeat sequences associated with hereditary neurological diseases. *J Biol Chem* 1998; **273**: 19532–19541.
9 Richard GF, Goellner GM, McMurray CT, Haber JE: Recombination-induced CAG trinucleotide repeat expansions in yeast involve the MRE11–RAD50–XRS2 complex. *EMBO J* 2000; **19**: 2381–2390.
10 Jakupciak JP, Wells RD: Gene conversion (recombination) mediates expansions of CTG · CAG repeats. *J Biol Chem* 2000; **275**: 40003–40013.
11 Djian P, Hancock JM, Chana HS: 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* 1996; **93**: 417–421.
12 Andrés AM, Soldevila M, Lao O *et al*: Comparative genetics of functional trinucleotide tandem repeats in humans and apes. *J Mol Evol* 2004; **59**: 329–339.
13 Goto J, Watanabe M, Ichikawa Y *et al*: Machado–Joseph disease gene products carrying different carboxyl termini. *Neurosci Res* 1997; **28**: 373–377.
14 Stevanin G, Lebre AS, Mathieux C *et al*: Linkage disequilibrium between the spinocerebellar ataxia 3/Machado–Joseph disease mutation and two intragenic polymorphisms, one of which, X359Y, affects the stop codon. *Am J Hum Genet* 1997; **60**: 1548–1552.
15 Costa MC, Sequeiros J, Maciel P: Identification of three novel polymorphisms in the MJD1 gene and study of their frequency in the Portuguese population. *J Hum Genet* 2002; **47**: 205–207.
16 Benson G: Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 1999; **27**: 573–580.
17 Gaspar C, Lopes-Cendes I, Hayes S *et al*: Ancestral origins of the Machado–Joseph disease mutation: a worldwide haplotype study. *Am J Hum Genet* 2001; **68**: 523–528.
18 Schneider S, Roessli D, Excoffier L: *Arlequin Ver. 2.000: A Software for Population Genetics Data Analysis*. Switzerland: Genetics and Biometry Laboratory, Department of Anthropology, University of Geneva, 2000.
19 Zhang J, Rowe WL, Clark AG, Buetow KH: Genomewide distribution of high-frequency, completely mismatching SNP haplotype pairs observed to be common across human populations. *Am J Hum Genet* 2003; **73**: 1073–1081.
20 Kong A, Gudbjartsson DF, Sainz J *et al*: A high-resolution recombination map of the human genome. *Nat Genet* 2002; **31**: 241–247.
21 Silveira I, Miranda C, Guimarães L *et al*: Trinucleotide repeats in 202 families with ataxia: a small expanded $(CAG)_n$ allele at the SCA17 locus. *Arch Neurol* 2002; **59**: 623–629.
22 Chattopadhyay B, Basu P, Gangopadhyay PK *et al*: Variation of CAG repeats and two intragenic polymorphisms at SCA3 locus among Machado–Joseph disease/SCA3 patients and diverse normal populations from eastern India. *Acta Neurol Scand* 2003; **108**: 407–414.
23 Takano H, Cancel G, Ikeuchi T *et al*: Close associations between prevalences of dominantly inherited spinocerebellar ataxias with CAG-repeat expansions and frequencies of large normal CAG alleles in Japanese and Caucasian populations. *Am J Hum Genet* 1998; **63**: 1060–1066.
24 Rubinsztein DC, Leggo J, Coetzee GA, Irvine RA, Buckley M, Ferguson-Smith MA: Sequence variation and size ranges of CAG repeats in the Machado–Joseph disease, spinocerebellar ataxia type 1 and androgen receptor genes. *Hum Mol Genet* 1995; **4**: 1585–1590.

25 Limprasert P, Nouri N, Heyman RA *et al*: 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* 1996; **5**: 207–213.
26 Azevedo L, Climent C, Vilarinho L, Calafell F, Amorim A: Evidence for mutational *cis*-acting factors affecting mutagenesis in the ornithine transcarbamylase gene. *Hum Mutat* 2004; **24**: 273.
27 Mittal U, Srivastava AK, Jain S, Jain S, Mukerji M: Founder haplotype for Machado–Joseph disease in the Indian population: novel insights from history and polymorphism studies. *Arch Neurol* 2005; **62**: 637–640.
28 Brunner HG, Jansen G, Nillesen W *et al*: Brief report: reverse mutation in myotonic dystrophy. *N Engl J Med* 1993; **328**: 476–480.
29 O'Hoy KL, Tsilfidis C, Mahadevan MS *et al*: Reduction in size of the myotonic dystrophy trinucleotide repeat mutation during transmission. *Science* 1993; **259**: 809–812.
30 Warner JP, Barron LH, Fitzpatrick DR, Brock DJH: A gene conversion event at the Huntington's CAG repeat. *Am J Hum Genet* 1996; **59**: 1697.
31 Labuda M, Labuda D, Miranda C *et al*: Unique origin and specific ethnic distribution of the Friedreich ataxia GAA expansion. *Neurology* 2000; **54**: 2322–2324.
32 Imbert G, Kretz C, Johnson K, Mandel JL: Origin of the expansion mutation in myotonic dystrophy. *Nat Genet* 1993; **4**: 72–76.
33 Sobczak K, Krzyzosiak WJ: Patterns of CAG repeat interruptions in SCA1 and SCA2 genes in relation to repeat instability. *Hum Mutat* 2004; **24**: 236–247.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)