RESEARCH NOTE

Evolutionary appearance of mononucleotide repeats in the coding sequences of four genes in primates

ARIANE PAOLONI-GIACOBINO[2*] and JOHN RICHARD CHAILLET[1]

[1]*Department of Molecular Genetics and Biochemistry, University of Pittsburgh, PA 15213, USA*
[2]*Department of Genetic Medicine and development, Geneva University Medical School 1, Michel-Servet 1211 Geneva 4, Switzerland*

Introduction

The replication instability of homopolymeric nucleotide runs is associated with several human diseases and has linked with colorectal cancers (Shibata *et al*. 1994). Little is known, however, about the evolutionary pattern of these sequences. The aim of this study was to analyse the long homopolymeric nucleotide runs that are unstable regions subject to replication errors, present in the coding sequence of four genes, *TGFBRII, BAX, MSH6* and *MLH3*, in six nonhuman primates encompassing 55 million years of evolution, and in humans. The results show that the $(A)_{10}$ of *TGFBRII* and the $(G)_8$ of *BAX* are already present in the ancestral ring-tailed lemur. The $(C)_8$ of *MSH6* appears during the orangutan–gorilla divergence and the $(A)_9$ of *MLH3* coexists with a putative precursor of the homopolymeric nucleotide runs until the rhesus macaque–orangutan divergence. Variations between clones of the homopolymeric nucleotide runs in each species are probably somatic changes due to slipped mispairing. They are not observed with homopolymeric nucleotide runs of less than eight nucleotides and, with runs of eight nucleotides or more, there is a tendency toward an increase in the sequence variations with the length of the run.

Materials and methods

The *TGFBRII* gene that codes for an inhibitor of cell proliferation contains a $(A)_{10}$ repeat. The *BAX* gene codes for an inhibitor of apoptosis and contains a $(G)_8$ repeat. *MSH6* and *MLH3* code for proteins involved in postreplicative DNA mismatch repair (MMR), and contain a $(C)_8$ and a $(A)_9$ repeat, respectively.

The phylogenetic primate DNA panel PRP00001 was obtained through the Coriell Cell Repositories (Corriel Institute for Medical Research, Camden, NJ 08103, USA). Genomic DNA from the following species was analysed: *Pan troglodytes* (chimpanzee, # NG06939), *Pan paniscus* (bonobo, # NG05253), *Gorilla gorilla* (gorilla, # NG05251), *Pongo pygmaeus* (Sumatra orangutan, # NG12256), *Macaca mulata* (rhesus macaque, # NG07109), and *Lemur catta* (ring-tailed lemur, # NG07099). The six nonhuman primate species encompass 55 million years (MY) of evolution. Great apes are represented by the chimpanzee, bonobo, gorilla and orangutan; old world monkeys by the rhesus macaque, and the ancestral prosimians by the ring-tailed lemur. Humans shared a common ancestor with the chimpanzee/bonobo, gorilla, and orangutan approximately 6, 8, and 16 MY ago (Chen and Li 2001), and with the rhesus macaque and ring-tailed lemur approximately 27–30 and 55 MY ago, respectively (Goodman 1999; Fukami-Kobayashi *et al*. 2005).

TGFBRII

The initial chimpanzee sequence was obtained by performing a blast search of human GenBank sequence NM_003242 against chimpanzee genome at http://www.genome.ucsc.edu/.

Amplifications of the *TGFBRII* gene fragment were conducted under the following reaction conditions: 100 ng of genomic DNA, 0.2 mM deoxynucleoside triphosphates (dNTPs), 0.4 μM of each primer, and 1.25 U of *Taq* DNA polymerase (Invitrogen) and corresponding 1X PCR Buffer (Invitrogen), 1.5 mM $MgCl_2$, in a 50 μl reaction mixture. Amplification of a 150-bp fragment with oligonucleotides designed on the corresponding chimpanzee sequence was performed with TGFBRF: 5′- aca cta gag aca gtt tgc ca -3′ and TGFBRR: 5′- gtc att gca ctc atc aga gc -3′. The PCR cycling conditions were as follows: 94°C for 5 min, 30 cycles of 94°C for 1 min, 58°C for 30 s, 72°C for 30 s and a final extension of 72°C for 10 min.

*For correspondence. E-mail: arine.giacobino@medecine.unige.ch.

Keywords. homopolymeric runs; slippage; frameshift; sequence evolution; primates.

BAX

The initial chimpanzee sequence was obtained by performing a blast search of human GenBank sequence NM_138764 against chimpanzee genome at http://www.genome.ucsc.edu/.

Amplifications of the *BAX* gene fragment were conducted under the following reaction conditions: 100 ng of genomic DNA, 0.2 mM deoxynucleoside triphosphates (dNTPs), 0.4 μM of each primer, and 1.25 U of *Taq* DNA polymerase (Invitrogen) and corresponding 1 × PCR Buffer (Invitrogen), 3 mM $MgCl_2$, in a 50 μl reaction mixture. Amplification of a 143-bp fragment with oligonucleotides designed on the corresponding chimpanzee sequence was performed with BAXF: 5′- ttc atc cag gat cga gca gg -3′ and BAXR: 5′- tgc agc tcc atg tta ctg tcc -3′. The PCR cycling conditions were as follows: 94°C for 5 min, 30 cycles of 94°C for 1 min, 57°C for 30 s, 72°C for 30 s and a final extension of 72°C for 10 min.

MSH6

The initial chimpanzee sequence was obtained by performing a blast search of human GenBank sequence NM_000179 against chimpanzee genome at http://www.genome.ucsc.edu/.

Amplifications of the *MSH6* gene fragment were conducted under the following reaction conditions: 100 ng of genomic DNA, 0.2 mM deoxynucleoside triphosphates (dNTPs), 0.4 μM of each primer, and 1.25 U of *Taq* DNA polymerase (Invitrogen) and corresponding 1 × PCR Buffer (Invitrogen), 3 mM $MgCl_2$, in a 50 μl reaction mixture. Amplification of a 145-bp fragment with oligonucleotides designed on the corresponding chimpanzee sequence was performed with MSHF: 5′- gat ggt cct atg tgt cgc c -3′ and MSHR: 5′- cct cac agc cta tta gaa tg -3′. The PCR cycling conditions were as follows: 94°C for 5 min, 30 cycles of 94°C for 1 min, 60°C for 30 s, 72°C for 30 s and a final extension of 72°C for 10 min.

MLH3

The initial chimpanzee sequence was obtained by performing a blast search of human GenBank sequence NM_014381 against chimpanzee genome at http://www.genome.ucsc.edu/.

Amplifications of the *MLH3* gene fragment were conducted under the following reaction conditions: 100 ng of genomic DNA, 0.2 mM deoxynucleoside triphosphates (dNTPs), 0.4 μM of each primer, and 1.25 U of *Taq* DNA polymerase (Invitrogen) and corresponding 1 × PCR Buffer (Invitrogen), 3 mM $MgCl_2$, in a 50 μl reaction mixture. Amplification of a 172-bp fragment with oligonucleotides designed on the corresponding chimpanzee sequence was performed with MLHF: 5′- gat gct act gaa gtg gga tgc -3′ and MLHR: 5′- cta cat gag tta taa agc ca -3′. The PCR cycling conditions were as follows: 94°C for 5 min, 30 cycles of 94°C for 1 min, 60°C for 30 s, 72°C for 30 s and a final extension of 72°C for 10 min.

Cloning and sequencing

PCR-amplified fragments were cloned into pCRII-TOPO vector (Invitrogen) through the TOPO-TA cloning system (Invitrogen). Purification of selected colonies was performed with the Quiaprep Spin Miniprep (Quiagen). Five positive clones with the insert of interest for each gene in each species were sequenced with universal oligonucleotides M13 Reverse on pCRII-TOPO vector. Sequences were determined using the Applied Biosystems Big Dye sequencing kit on an ABI 3700 automated sequencer (Applied Biosystems).

Control for Taq DNA polymerase errors

As a control for *Taq* polymerase errors, for each one of the four genes of interest, a cloned human fragment with the homopolymeric run of original length was PCR amplified with the same oligonucleotides and conditions as those used for the primate amplification of the corresponding fragment. Then the fragments were subcloned into pCRII-TOPO vector and 10 clones for each gene of interest were analysed by sequencing with oligonucleotide M13 Reverse.

Results

Table 1 shows the homopolymeric nucleotide run sequences obtained for our target genes in humans and the six nonhuman primate species studied. As described earlier, we excluded the possibility that observed variations were due to *Taq* polymerase errors during PCR.

Table 1. The different stretch length obtained for our four genes of interest, *TGFBRII, BAX, MSH6 and MLH3.* Five clones were analysed for each gene and each of the seven species, human, chimpanzee, bonobo, gorilla, orangutan, rhesus macaque and ring-tailed lemur. The evolutionary distances are indicated in million years ago: GA, great apes; OW, old world monkeys; P, prosimians. In the presentation of the data, the number of time each stretch of a particular length has been found is indicated.

Genes	Human	Chimpanzee	Bonobo	Gorilla	Orangutan	Rh. Macaque	R-tailed lemur
TGFBRII$(A)_{10}$	$2\times(A)_{10}$:$3\times(A)_9$	$3\times(A)_{10}$:$2\times(A)_9$	$5\times(A)_{10}$	$4\times(A)_{10}$:$1\times(A)_9$	$4\times(A)_{10}$:$1\times(A)_9$	$4\times(A)_{10}$:$1\times(A)_9$	$5\times(A)_{10}$
BAX$(G)_8$	$5\times(G)_8$	$5\times(G)_6$	$3\times(G)_8$:$1\times(G)_7$:$1\times(G)_6$	$4\times(G)_8$:$1\times(G)_7$	$5\times(G)_8$	$5\times(G)_8$	$5\times(G)_8$
MSH6$(C)_8$	$5\times(G)_8$	$5\times(G)_8$	$5\times(G)_8$	$4\times(C)_8$:$1\times(C)_7$	$5\times(C)_3G(C)_4$	$5\times((C)_4T(C)_3)$	$5\times((CTGT)(C)_4)$
MLH3$(A)_9$	$5\times(A)_9$	$5\times(A)_9$	$5\times(A)_9$	$5\times(A)_9$	$5\times(A)_9$	$4\times(A)_9$:$1\times((A)_4G(A)_7)$	$4\times(A)_9$:$1\times((A)_4G(A)_7)$

It can be seen that the $(A)_{10}$ of *TGFBRII* and the $(G)_8$ of *BAX* are already present in the ancestral ring-tailed lemur. The $(C)_8$ of *MSH6*, however appears only in the gorilla. In the ring-tailed lemur, the corresponding sequence exists as a complex CTGTCCCC sequence, that evolves into two small stretches of $(C)_4$ and $(C)_3$, separated by a T or a G in the rhesus macaque and orangutan, respectively. Then, the $(C)_8$ appears during the orangutan–gorilla divergence, possibly by a G-C substitution. The $(A)_9$ of *MLH3* is already present in the ancestral ring-tailed lemur. However, in the ring-tailed lemur and rhesus macaque, this $(A)_9$ coexists with two stretches of $(C)_4$ and $(C)_7$ separated by a G that could be considered an evolutionary remnant this remnant transforms into $(A)_9$ during the rhesus macaque–orangutan divergence, possibly by a GAA deletion.

It is noteworthy that the long homopolymeric nucleotide runs of the four genes analysed in this study have persisted without any variations, from their presence in ring-tailed lemurs up through the evolution of humans. The variations between clones of the long homopolymeric nucleotide run sequences in each of the species analysed is probably the result of slipped mispairing. No variation is observed in the short $(C)_3$ and $(C)_4$ runs of *MSH6* in the rhesus macaque and in the orangutan. Similarly, no variation is observed in the $(A)_4$ and $(A)_7$ runs of *MLH3* in the ring-tailed lemur and in the rhesus macaque. In the $(C)_8$ of *MSH6*, however, one $(C)_7$ is detected in the 20 clones sequenced from gorillas to humans (5% variation). In the $(G)_8$ of *BAX*, two $(G)_7$ and one $(G)_6$ are detected in the 35 clones sequenced from the ring-tailed lemur to humans (6% variation). In the $(A)_9$ of *MLH3*, there is no variation from orangutans to the humans, but in the $(A)_{10}$ of *TGFBRII*, eight $(A)_9$ are detected in the 35 clones sequenced from the ring-tailed lemur to humans (23% variation). These results show no slipped mispairing with homopolymeric runs of less than eight nucleotides. With homopolymeric runs of eight nucleotides or more, there is a tendency toward an increase in the frequency of the variations with the length of the run. One exception is the remarkable stability of the $(A)_9$ of *MLH3* gene.

Figure 1 shows the alignments of 143-172-bp long segments of our genes of interest surrounding the homopolymeric nucleotide runs. The degrees of homology observed when comparing human sequences to all the six nonhuman primate species of this study are 89.19% for *TGFBRII*, 90.2% for *BAX*, 89.63% for *MSH6* and 87.22% for *MLH3*. $A(T)_7$ sequence is observed in *MSH6*, 44 nucleotides downstream of the 3′ end of the homopolymeric nucleotide run, and this is totally conserved across the seven species analysed.

Discussion

The variations between clones of the long homopolymeric nucleotide run sequences probably reflect somatic variations. The $(T)_7$ sequence observed in *MSH6*, was found to be totally conserved in the seven species analysed. Indeed, tracts less than eight nucleotides long seem resistant to slippage, suggesting that there is a critical size above which a homopolymeric tract can be subjected to slippage (Dechering *et al*. 1998). However, computer simulations using analysis of genomic sequences in nine species suggested that short repeats, such as $(C)_4$, might have a susceptibility to insertions and deletions 10-15-fold higher than nonrepetitive sequences (Dieringer and Schlotterer 2003). The present study shows that with homopolymeric runs of eight nucleotides or more, there is a tendency toward an increase in the frequency of the variations with the length of the run. These observations confirm previous studies (Paoloni-Giacobino *et al*. 2001, 2002) and extend them to nonhuman primates.

The homopolymeric nucleotide runs of *TGFBRII* and of *BAX* were already present in the ancestral ring-tailed lemur, whereas those of the MMR genes *MSH6* and *MLH3* arose during the orangutan–gorilla and rhesus macaque–orangutan divergences, respectively. The present study also suggests a mechanism for the formation, during evolution, of the slippage-prone runs. It shows that the homopolymeric nucleotide runs are not formed by extension of a smaller run, but rather by change in a hetero–nucleotide separating two homopolymeric stretches of less than eight nucleotides. Two examples of this evolution are the emergence of the $(C)_8$ of *MSH6* during the orangutan–gorilla divergence from a $(C)_3G(C)_4$ and the emergence of the $(A)_9$ of *MLH3* during the rhesus macaque–orangutan divergence from a $(A)_4G(A)_7$. This type of sequence might, thus, represent the evolutionary precursors of homopolymeric nucleotide runs.

The deleterious effect of slipped mispairing in the coding sequence of genes is a possible truncation of the resulting protein by change in the reading frame. The coding regions represent an essentially different environment to the repeats than the noncoding regions. The main difference is that evolutionary constraints imposed on coding regions by protein function are greater than on noncoding regions (Borstnik and Pumpernik 2002). The homopolymeric run appearance seems unbeneficial from an evolutionary point of view. The driving force of the expansion of homopolymeric runs might be the polymerase slippage process. It has been proposed that slipped strand replication is a major force in the evolution of genes and genomes (Dechering *et al*. 1998), and it has been implicated in a large number of human genetic diseases (Lavoie *et al*. 2003).

GenBank Accession Numbers

The following accession numbers have been provisionally applied to the primate sequences described: DQ229079, DQ229080, DQ229081, DQ229082, DQ229083, DQ229084, DQ229085, DQ229086, DQ229087, DQ229088, DQ229089, DQ229090, DQ229091, DQ229092, DQ229093, DQ229094, DQ229095, DQ229096, DQ229097, DQ229098.

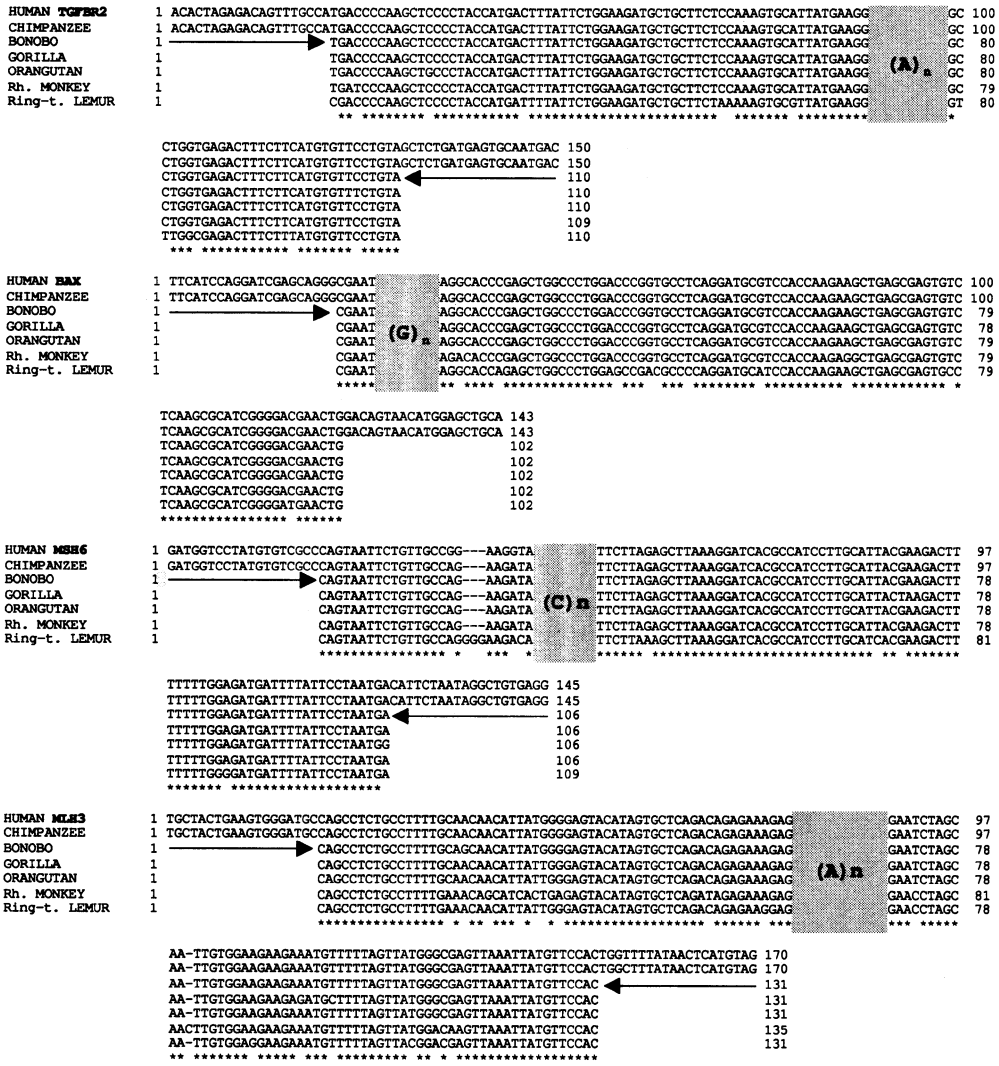

Figure 1. Triplet coding frame alignment of the sequences obtained from the human, chimpanzee, bonobo, gorilla, orangutan, rhesus monkey and ring-tailed lemur, for the four genes of interest, *TGFBR2*, *BAX*, *MSH6*, *MLH3*. Stars correspond to the nucleotides conserved in all five species. Arrows indicate the primers designed on the human and identical portion of chimpanzee sequence. Grey squares cover the stretch of interest in each gene, and the nucleotide in cause for each. The 'n' in subscript indicates that the stretch is of variable length, as detailed in table 1.

Acknowledgements

This work was funded by grants from the Fondation Suisse des Bourses Médecine et Biologie (to APG), and by the National Institutes of Health, USA (to JRC).

References

Borstnik B. and Pumpernik D. 2002 Tandem repeats in protein coding regions of primate genes. *Genome Res.* **12**, 909–915.

Chen F. C. and Li W. H. 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.

Dechering K. J., Cuelenaere K., Konings R. N. and Leunissen J. A. 1998 Distinct frequency-distributions of homopolymeric DNA tracts in different genomes. *Nuclic Acids Res.* **26**, 4056–4062.

Dieringer D. and Schlotterer C. 2003 Two distinct modes of microsatellite mutation processes: evidence from the complete genomic sequences of nine species *Genome Res.* **13**, 2242–2251.

Fukami-Kobayashi K., Shiina T., Anzai T., Sano K., Yamazaki M., Inoko H. and Tateno Y. 2005 Genomic evolution of MHC class I region in primates. *Proc. Natl. Acad. Sci. USA* **102**, 9230–9234.

Goodman M. 1999 The genomic record of humankind's evolutionary roots. *Am. J. Hum. Genet.* **64**, 31–39.

Lavoie H., Debeane F., Trinh Q. D., Turcotte J. F., Corbeil-Girard L. P., Dicaire M. J. et al. 2003 Polymorphism, shared functions and convergent evolution of genes with sequences coding for polyalanine domains. *Hum. Mol. Genet.* **12**, 2967–2979.

Paoloni-Giacobino A., Rey-Berthod C., Couturier A., Antonarakis S. E. and Hutter P. 2002 Differential rates of frameshift alterations in four repeat sequences of hereditary nonpolyposis col-

orectal cancer tumors. *Hum. Genet.* **111**, 284–289.

Paoloni-Giacobino A., Rossier C., Papasavvas M. P. and Antonarakis S. E. 2001 Frequency of replication/transcription errors in (A)/(T) runs of human genes. *Hum. Genet.* **109**, 40–47.

Shibata D., Peinado M. A., Ionov Y., Malkhosyan S. and Perucho M. 1994 Genomic instability in repeated sequences is an early somatic event in colorectal tumorigenesis that persists after transformation. *Nat. Genet.* **6(3)**, 273–281.

Received 15 February 2006