Journal of Genetics, Vol. 97, No. 5, December 2018, pp. 1473–1478
https://doi.org/10.1007/s12041-018-1039-0

RESEARCH NOTE

Gene fusion of heterophyletic gamma-globin genes in platyrrhine primates

JOSÉ IGNACIO ARROYO[1,2,3]* and MARIANA F. NERY[4]

[1] *Departamento de Ecología, Facultad de Ciencias Biológicas, Pontificia Universidad Católica de Chile, Santiago 8331150, Chile*
[2] *Instituto de Ecología y Biodiversidad (IEB), Santiago 7800003, Chile*
[3] *Center for Climate and Resilience Research (CR)2, Santiago 8370449, Chile*
[4] *Departamento de Genética, Evolução, Microbiologia e Imunologia, Instituto de Biologia, Universidade Estadual de Campinas, Campinas 13148-252, Brazil*
*For correspondence. E-mail: jiarroyo@uc.cl.

Received 19 October 2017; revised 24 March 2018; accepted 8 August 2018; published online 23 November 2018

Abstract. We performed phylogenetic analyses of HBG genes to assess its origin and interspecific variation among primates. Our analyses showed variation in HBG genes copy number ranging from one to three, some of them pseudogenes. For platyrrhines HBG genes, phylogenetic reconstructions of flanking regions recovered orthologous clades with distinct topologies for 5′ and 3′ flanking regions. The 5′ region originated in the common ancestor of platyrrhines but the 3′ region had an anthropoid origin. We hypothesize that the platyrrhine HBG genes of 5′ and 3′ heterophyletic origins arose from subsequent fusions of the (earlier) platyrrhine 5′ portion and the (later) anthropoid 3′ portion.

Keywords. gene duplication; gene fusion; gene conversion; recombination; globins; primates.

Introduction

Gene duplication and fusion are common mechanisms of genomic innovation (Wen and Irwin 1999; Lynch and Conery 2003; Nei and Rooney 2005; Gaudry *et al*. 2014). The beta globin gene cluster is a clear example of this as these genes have undergone multiple recombination events and has been reinvented during its evolutionary history. These tandemly arranged genes make up the beta subunits of haemoglobin protein, which are differentially expressed during development according to their positions in the gene cluster (Sankaran *et al*. 2010). In placental mammals, the genes localized in the 5′ portion of the cluster (HBE–HBG–HBH) are expressed early and the genes localized in the 3′ portion (HBD–HBB) are expressed late in development. The copy number of this gene cluster varies among species as a result of independent duplications and/or differential retentions after duplications. For instance while Laurasiatheria have lost the HBG gene while retaining the HBH gene, Atlantogeneta and Euarchontoglires have lost the HBH gene while retaining the HBG gene (Opazo *et al*. 2008). Within Euarchontoglires, Anthropoid primates possess duplicated copies of the gamma globin (HBG) gene, however its evolutionary origin has not been rigorously established (see supplementary material in Fitch *et al*. 1991).

Methods

Genomic structure characterization and phylogenetic analysis

We obtained the genomic sequence of the beta-globin gene cluster (HBE–HBGs–HBH–HBD–HBB) plus 1-kb upstream flanking and downstream flanking sequences for 26 primate species from the Whole Genome Shotgun (WGS) database (table 1 in electronic supplementary material at http://www.ias.ac.in/jgenet/). Our survey included data for representative species of the main groups

Electronic supplementary material: The online version of this article (https://doi.org/10.1007/s12041-018-1039-0) contains supplementary material, which is available to authorized users.

Table 1. Gene repertoire of the beta-blogin gene cluster among primates and non-primates.

Group	HBE	HBG	HBH	HBD	HBB
OWM	1	1, 2, 1-2	p	1	1
Apes		1-2, 1_p-2, 2, 1-2_p-3_p			
Platyrrhines (or NWM)	1	1-2, 2	p	1	1
NWM (Atelidae)	1	1_p-2	p	1	1
Nonanthropoid primates	1	1	p	0	1
Nonprimates	1	1	0	1	1

Numbers represent the copy number found in the beta-globin gene cluster whereas the subindex 'p' denotes a pseudogene (understood as a partial sequence). Distinct numbers separated by commas indicate the alternative repertoires found among species. For instance, OWM have species with the -T1 copy, the -T2 or both -T1 and -T2. Among OWMs, only *Colobus* and *P. anubis* have two copies but all the remaining species had one; species of the *Macaca* genus have the -T2 and *Chlorocebus* has the -T1 copy. In Apes, *Homo*, *Pongo* and *Nomascus* have two copies, *P. troglodytes* has three (two partial and one complete) and *Gorilla* retained only a partial -T1 (3rd exon) and a complete -T2. Among platyrrhines (NWMs), *Aotus* has only a -T2 copy and in Atelidae only the -T2 copy is complete and the 3rd exon of -T1 copy was deleted (as previously reported; Chiu *et al.* 1999). Nonanthropoids as well as Dermoptera and Scadentia species had one HBG copy. See details in figure 1 electronic supplementary material.

including catarrhines (Old World monkeys (OWMs) and apes) and its sister clade, the platyrrhines (also called New World monkeys (NWMs)) both conforming the anthropoids as well as nonanthropoids (tarsiers and strepsirrhines, figure 1a; figure 1 in electronic supplementary material). To annotate beta-globin sequences, we used well-annotated reference genes from Ensembl database and aligned them against the whole piece using Blas2seq 2.2 (Tatusova and Madden 1999). Initially, each beta-like globin sequence was annotated with the genus name followed by a T and a number indicating their position in the gene cluster. Further comparisons of genomic sequences (e.g. to search for breakpoints) were performed using dot plots as implemented in PipMaker (Schwartz *et al.* 2010). For phylogenetic reconstruction, we aligned all beta-globin sequences including 1-kb upstream flanking and downstream flanking sequences using multiple alignment using fast fourier transform (MAFFT; G-INS-i strategy; Katoh *et al.* 2009). We inferred phylogenetic relationships using maximum likelihood (ML) and Bayesian approaches as implemented in iqTree (Trifinopoulos *et al.* 2016) and Mr. Bayes v3.2.2 (Ronquist *et al.* 2012), respectively. For ML, we used a GTR+G model and support for the nodes were estimated with 1000 ultrafast bootstrap pseudoreplicates. For Bayesian analyses, two simultaneous independent runs were performed for 1×10^6 iterations of a Markov chain Monte Carlo algorithm, with six simultaneous chains sampling trees every 1000 generations. Support for the nodes and parameter estimates were derived from a majority rule consensus of the last 50% of the trees sampled after convergence. The average standard deviation of split frequencies remained 0.01 after the burn-in threshold. Different trees were reconstructed for alignments based on 1-kb upstream flanking, exons 1, 2 and 3, introns 2 (as is larger than intron 1 and further has more data) and 1-kb downstream flanking sequences.

Results and discussion

Genomic structure and phylogenetic reconstruction of beta-globin gene cluster in primates

Among the 26 primate species surveyed, we found the whole beta-globin gene cluster for 17 and for the other nine species we found beta-globin genes in separate genomic pieces (contigs and/or scaffolds) (table 1 in electronic supplementary material). All surveyed species have a single HBE gene at the 5′ end of the cluster, a single-copy HBH pseudogene, a HBD and a HBB at the 3′ end. The HBG gene is found in strepshrirrines and tarsiers as a single copy, and anthropoids have two duplicated divergent copies (as previously reported for some species (Fitch *et al.* 1990, 1991) (table 1; figure 1 in electronic supplementary material). In our survey, we found that not all anthropoids have two complete copies of HBG genes. Apes had the greatest variation in their repertoire followed by OWMs and platyrrhines which had in most cases two copies, except in *Aotus* which has only a -T2 copy and in Atelidae where 3rd exon of -T1 copy was deleted and only the -T2 copy is complete (as previously reported; Chiu *et al.* 1999) (table 1; figure 1 in electronic supplementary material). Nonanthropoids as well as Dermoptera and Scadentia species had one HBG copy (table 1; figure 1 in electronic supplementary material). Previous studies have characterized the genomic structure of beta-globin gene cluster in primates, but only

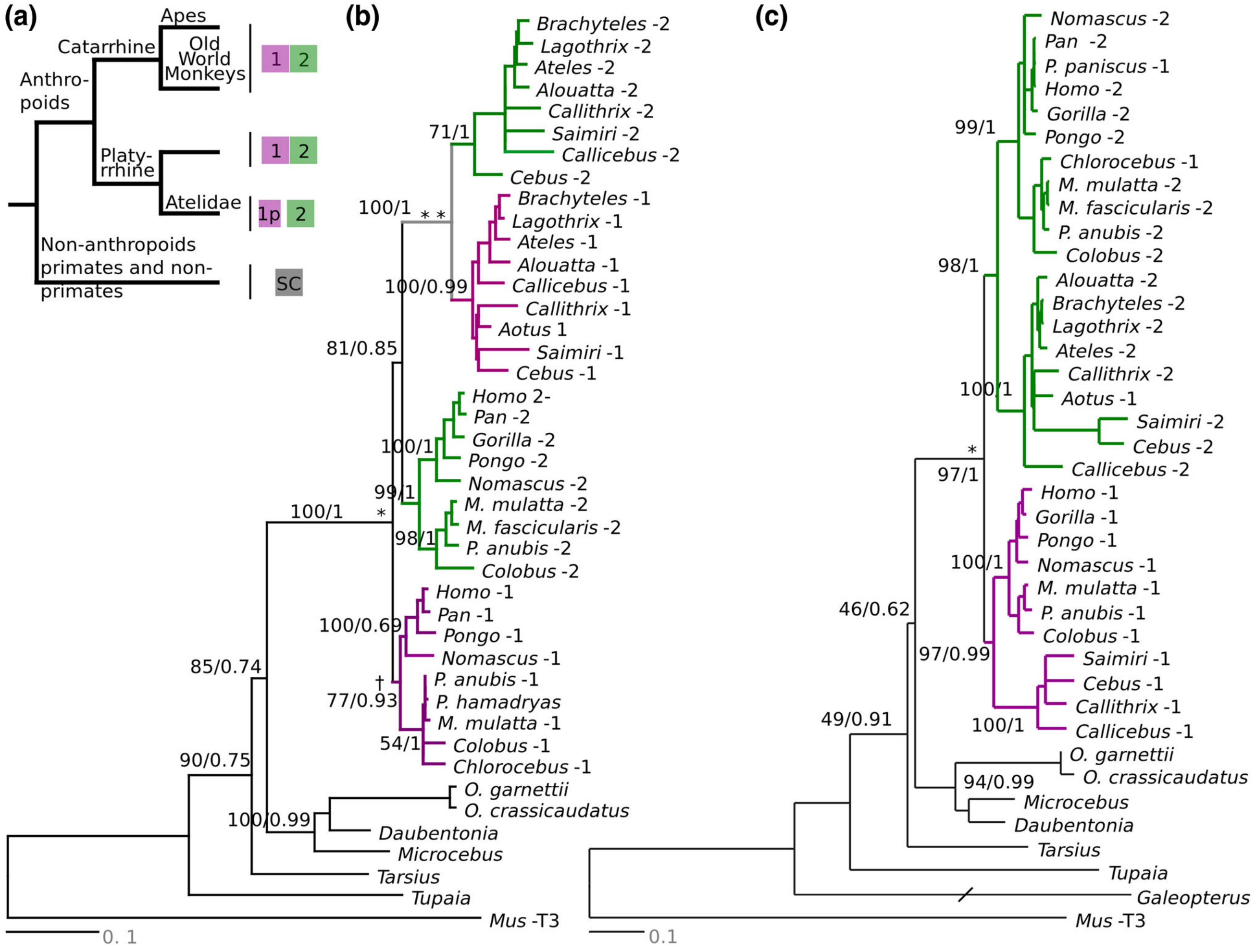

Figure 1. (a) Tree of primate groups included in this study. Violet and green colours in boxes and the respective terminal branches of phylogenetic trees denote the -T1 and -T2 copies, respectively. Nonanthropoid primates and (most) nonprimates have a single copy (SC), within anthropoids all have two complete copies except in Atelidae where the second copy have been pseudogenized (p). Phylogenetic trees for HBG genes based on (b) 5′ flanking, (c) 3′ flanking sequences. Tree was rooted with *Mus*. Supports on the nodes are bootstrap and Bayesian posterior probabilities. Suffix 1 and 2 after names of genera indicate HBG-T1 and HBG-T2, respectively. The 5′ flanking-based tree shows a first duplication in anthropoids (*) followed by a subsequent loss (†) of this ancient duplicate in platyrrhines and a new (earlier) duplication (**) derived from the -T2 copy that occurred in platyrrhines. The 3′ flanking-based tree shows a single origin in the common ancestor of anthropoids (the slash indicate that the branch is not at scale).

for a restricted number of species (Tagle *et al*. 1988; Fitch *et al*. 1990, 1991; Hayasaka *et al*. 1993; Chiu *et al*. 1996). This new data shows that the evolution of these genes is not so conservative as previously thought. Phylogenetic analyses based on the three exons of all beta-like globin genes recovered the monophyly of each of the beta globin paralogous (figure 2 in electronic supplementary material). Within the HBG clade there were recovered paralogous groups of duplicated genes, which could be attributable to recent duplication or conversion (Nei and Rooney 2005).

Phylogenetic reconstruction using exon and intron sequences of HBG genes

To further investigate whether the recovery of paralogous groups is due to recent gene duplication or conversion, we performed phylogenetic analysis separately for each exon and intron. If a recent duplication occurred, we should recover paralogous groups in each of the different exon/intron segments. If it is due to conversion, because it often involves short DNA segments (as previously reported for primate beta-globin genes, Fitch *et al*. 1990), we should recover the nonconverted segments in orthologous groups but the converted in paralogous groups. Phylogenies reconstructed separately with one each of the three exon sequences recovered both paralogous and orthologous groups, indicating that different regions of the HBG genes evolved concertedly and divergently (the so-called mosaic evolutionary process, Wen and Irwin 1999; figure 3a in electronic supplementary material). For instance, the HBG sequences of exon 1 of *Homo* are recovered together but in exon 3 are recovered separately (figure 3a in electronic supplementary material). In other cases (such as in exons 1 and 3 of *Callithrix*; figure 3a in electronic supplementary material), paralogous groups are recovered together in each of the segments but in one case their branch lengths are equal and in the other they differ, denoting different rates of substitution. Phylogenetic reconstruction that were based on 1st and 2nd intron sequences (figure 3b in electronic supplementary material) showed similar mosaic patterns for species such as *Homo* and *Cebus* (figure 3b in electronic supplementary material).

Phylogenetic reconstruction using flanking sequences of HBG genes

Previous to phylogenetic reconstruction, to identify flanking sequences with a similarity of 100% (presumably due to conversion), we compared the sequence similarity between the 5′ region adjacent to the first exon and the 3′ region adjacent to the third exon in each of the paralogous pairs of the species surveyed. For the 5′ region, the species with the longest converted 5′ region was human (275 bp from first exon) and for the 3′ region the longest converted region was *Nomascus* (106 bp from third exon; table 2 in electronic supplementary material). In our alignment, we discarded the homologous segment within the 275 bp in the 5′ region and within the 106 bp in the 3′ region. In our phylogenies based on (nonconverted) flanking regions, we recovered distinct duplicative histories for 5′ and 3′ regions of HBG genes. The tree based on 5′ region depicts a first duplication that occurred in the common ancestor of anthropoids but was lost in platyrrhines and a second duplication occurred in this group (figure 1b). The tree based on the 3′ region is consistent with an anthropoid origin (as previously suggested, Fitch *et al.* 1991; figure 1c). To further test the contrasting topologies of 5′ and 3′ flanking sequences, we performed topology tests (as implemented in iqTree; Trifinopoulos *et al.* 2016). Our analysis rejected a single duplication in the common ancestor of anthropoids for a tree based on 5′ flanking sequences and also rejected a sister relationship for platyrrhine HBG genes for a tree based on 3′ flanking sequences (table 3 in electronic supplementary material). This reaffirms that the 5′ region of HBG of platyrrhines derived from a duplication that occurred in the common ancestor of this group and the 3′ region derives from a duplication that occurred in the common ancestor of anthropoids.

Recombinational signatures of gamma-globin genes in platyrrhine primates

As showed by phylogenetic reconstructions of 5′ and 3′ flanking genes, the upstream and downstream regions of HBG in platyrrhine primates have heterophyletic origins, which implies the occurrence of a gene fusion. If a recombination had occurred in the middle of the HBG, it would be possible that a signature of a recombination (a breakpoint in sequence similarity) could have persisted in extant species. Hence, for this reason, we looked for possible breakpoints in sequence similarity free of gene conversion in exons or introns in dot plots of the HBG-T1 and HBG-T2 paralogous. Among the species examined, only two showed a breakpoint: *Cebus* has a breakpoint in 3rd exon, and *Callithrix* in 2nd intron (figure 4 in electronic supplementary material). A breakpoint in sequence similarity is a common feature of recombination as observed in beta-globin genes of mammals (e.g. Neumann *et al.* 2010). For instance, platyrrhine primates of the family Atelidae have lost part of the second intron and 3rd exon of one of the duplicated HBG genes due to a recombination event in the 2nd intron (Chiu *et al.* 1996), which is evident when comparing HBG-1 and T2 (figure 4 in electronic supplementary material). Also, recent studies of copy number variation in human populations reported recurrent recombination between HBG paralogous genes, with breakpoint in different regions of HBG but mostly in second intron (Neumann *et al.* 2010).

Taking together our comparative and phylogenetic evidence of a breakpoint in 2nd intron and contrasting topologies of flanking-based phylogenetic reconstructions for 5′ (platyrrhine origin of HBGs) and 3′ (anthropoid origin of HBGs), we hypothesize an evolutionary pathway for the duplicated fusion genes of distinct evolutionary origin. From a molecular mechanistic standpoint, these fusion genes could have arisen by subsequent duplications and fusions. A first duplication occurred in the common ancestor of anthropoids and a second duplication in the common ancestor of platyrrhines, which was followed by an intragene recombination that involved the fusion of the 5′ region of platyrrhine origin and the 3′ region of anthropoid origin (figure 2). To test the role of positive selection in the anthropoid and platyrrhine duplications, we performed branch-site tests (see electronic supplementary material). We found that a significant positive selection occurred in the first exon in the anthropoid group, with one site inferred under this regime (see supplementary method and table 4 in electronic supplementary material for further details).

In summary, we have provided phylogenetic evidence that the upstream and downstream regions of the HBG genes of platyrrhines have an heterophyletic origin and have been subject of a mosaic molecular evolutionary process, where different regions have evolved divergently and concertedly or at different rates. Our study also constitutes additional evidence that support the hypothesis of elevated rates of gene gain and loss in primates (Hahn *et al.* 2007), a process that could have been driven neutrally by the increase in body size and the correlated reduction in effective population sizes which subsequently could have led to this accelerated duplication rates (Lynch and Conery 2003). An alternative possibility is that these changes are adaptive, but because gene conversion in coding regions erases the evolutionary history, it would be difficult to test

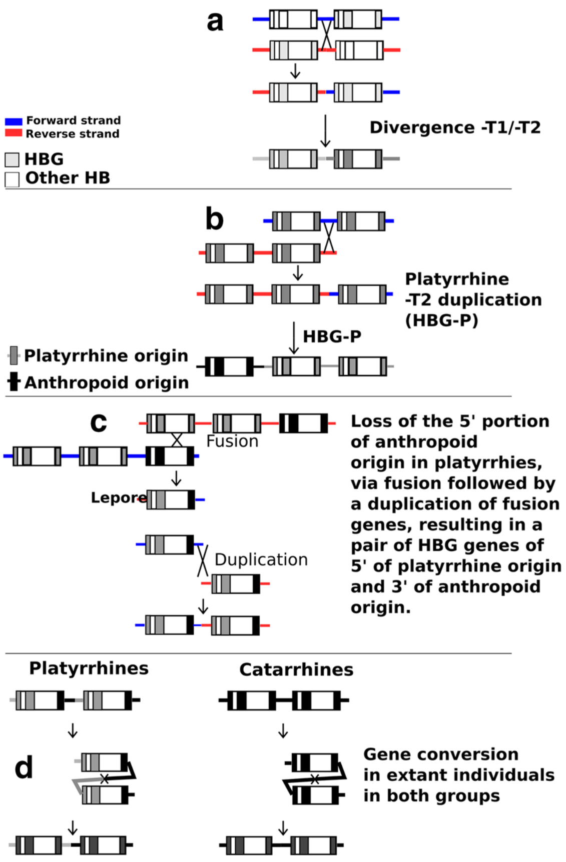
a
Forward strand
Reverse strand
HBG
Other HB
Divergence -T1/-T2
b
Platyrrhine
-T2 duplication
(HBG-P)
HBG-P
Platyrrhine origin
Anthropoid origin
c
Fusion
Loss of the 5' portion of anthropoid origin in platyrrhies, via fusion followed by a duplication of fusion genes, resulting in a pair of HBG genes of 5' of platyrrhine origin and 3' of anthropoid origin.
Lepore
Duplication
Platyrrhines
Catarrhines
d
Gene conversion in extant individuals in both groups

Figure 2. A model for the recombinational history of gamma–globin genes in primates. Our phylogenetic reconstructions for 5′ and 3′ flanking regions indicate that in addition to the duplication that originally generated duplicated HBG genes in anthropoid primates, platyrrhines underwent an additional duplication followed by subsequent fusions that resulted in the loss of the 5′ of anthropoid origin and their replacement by the earlier 5′ of platyrrhine origin. However, as shown by phylogenetic reconstructions of exons/introns, this history has been erased by concerted evolutionary process (presumably by gene conversion) in extant individuals. We hypothesize a possible recombinational history on four steps. (Here chromosome region is represented as a line and HBG genes are represented as boxes where the shaded regions are exons and nonshaded regions are introns. Crossing over is represented as a cross and their result with a short arrow. Long-term evolutionary events (divergence) are represented by a long arrow. Forward DNA strand is depicted in blue and reverse in red.) (a) A first duplication followed by divergence of an ancestral HBG occurred in the common ancestor (CA) of anthropoids leading to duplicated HBG. (b) A second duplication followed by divergence in the CA of platyrrhines led to platyrrhine duplicated HBGs. (c) A fusion of a 5′ portion of platyrrhine origin and a 3′ portion of anthropoid origin followed by a duplication among two fusion genes gave rise to duplicated heterophyletic HBGs. (d) In extant species gene conversion in exon–intron regions erased this duplicative history. Despite many possible recombinational pathways could have originated these pair of platyrrhine HBG heterophyletic genes, here we depicted a simple route that could explain the inferred histories of 5′ flanking tree (that 5′ region of HBGs have and origin in the CA of platyrrhines) and that of 3′ flanking 3′ (that 3′ region originated in the CA of anthropoids).

the effect of positive selection in these coding regions. Otherwise, it has been showed that the gamma-globin genes in platyrrhines have different expression patterns during development from that of catarrhines (Johnson *et al*. 2002; Sankaran *et al*. 2010). It is also possible that this process likely driven by innovations within upstream regulatory sequences could have been associated with these inferred gene duplications and fusions in platyrrhines. Accordingly, new comparative studies of flanking regulatory sequences could shed light on potential differential selective regimes in these two primate groups.

Acknowledgements

JIA acknowledges a Comisión Nacional de Investigación Científica y Tecnológica (CONICYT) National Doctoral Scholarship (no. 21130515).

References

Chiu C. H., Gregoire L., Gumucio D. L., Muniz J. A., Lancaster W. D. and Goodman M. 1999 Model for the fetal recruitment of simian gamma-globin genes based on findings from two New World monkeys *Cebus apella* and *Callithrix jacchus* (Platyrrhini, Primates). *J. Exp. Zool.* **285**, 27–40.

Chiu C. H., Schneider H., Schneider M. P., Sampaio I., Meireles C., Slighton J. L. *et al.* 1996 Reduction of two functional gamma-globin genes to one: an evolutionary trend in New World monkeys (infraorder Platyrrhini). *Proc. Natl. Acad. Sci. USA* **93**, 6510–6515.

Fitch D. H., Mainone C., Goodman M. and Slighton J. L. 1990 Molecular history of gene conversions in the primate fetal gamma-globin genes. Nucleotide sequences from the common gibbon, Hylobates lar. *J. Biol. Chem.* **265**, 781–793.

Fitch D. H., Bailey W. J., Tagle D. A., Goodman M., Sieu L. and Slighton J. L. 1991 Duplication of the gamma-globin gene mediated by L1 long interspersed repetitive elements in an early ancestor of simian primates. *Proc. Natl. Acad. Sci. USA* **88**, 7396–7400.

Gaudry M. J., Storz J. F., Butts G. T., Campbell K. L. and Hoffmann F. G. 2014 Repeated evolution of chimeric fusion genes in the beta-globin gene family of Laurasiatherian mammals. *Genome Biol. Evol.* **6**, 1219–1233.

Hayasaka K., Skinner C. G., Goodman M. and Slightom J. L. 1993 The gamma-globin genes and their flanking sequences in primates: findings with nucleotide sequences of capuchin monkey and tarsier. *Genomics* **18**, 20–28.

Johnson R. M., Gumucio D. and Goodman M. 2002 Globin gene switching in primates. *Comp. Biochem. Physiol. A. Mol. Integr. Physiol.* **133**, 877–883.

Katoh K., Asimenos G. and Toh H. 2009 Multiple alignment of DNA sequences with MAFFT. *Methods Mol. Biol.* **537**, 39–64.

Lynch M. and Conery J. S. 2003 The origins of genome complexity. *Science* **302**, 1401–1404.

Moleirinho A., Lopes A. M., Seixas S., Morales-Hojas R., Prata M. J. and Amorim A. 2015 Distinctive patterns of evolution of the δ-globin gene (HBD) in primates. *PLoS One* **10**, e0123365.

Nei M. and Rooney A. P. 2005 Concerted and birth-and-death evolution of multigene families. *Annu. Rev. Genet.* **39**, 121–152.

Neumann R., Lawson V. E. and Jeffreys A. J. 2010 Dynamics and processes of copy number instability in human γ-globin genes. *Proc. Natl. Acad. Sci. USA* **107**, 8304–8309.

Opazo J. C., Hoffmann F. G. and Storz J. F. 2008 Differential loss of embryonic globin genes during the radiation of placental mammals. *Proc. Natl. Acad. Sci. USA* **105**, 12950–12955.

Perelman P., Johnson W. E., Roos C., Seuánez H. N., Horvath J. E., Moreira M. A. *et al*. 2011 A molecular phylogeny of living primates. *PLoS Genet.* **7**, e1001342.

Ronquist F., Teslenko M., van der Mark P., Ayres D. L., Darling A., Höhna S. *et al.* 2012 MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539–542.

Sankaran V. G., Xu J. and Orkin S. H. 2010 Advances in the understanding of haemoglobin switching. *Br. J. Haematol.* **149**, 181–194.

Schwartz S., Zhang Z., Frazer K. A., Smit A., Riemer C., Bouck J. *et al*. 2000 PipMaker–a web server for aligning two genomic DNA sequences. *Genome Res.* **10**, 577–586.

Tagle D. A., Koop B. F., Goodman M., Slightom J. L., Hess D. L. and Jones R. T. 1988 Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J. Mol. Biol.* **203**, 439–455.

Tatusova T. and Madden T. 1999 BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *Fems Microbiol. Lett.* **174**, 247.

Trifinopoulos J., Nguyen L. T., von Haeseler A. and Minh B. Q. 2016 W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Res.* **44**, W232–W235.

Wen Y. and Irwin D. M. 1999 Mosaic evolution of ruminant stomach lysozyme genes. *Mol. Phylogenet. Evol.* **13**, 474–482.

Corresponding editor: N. G. Prasad