Heredity (2014) 113, 64–73

www.nature.com/hdy

ORIGINAL ARTICLE

Relaxed functional constraints on triplicate α-globin gene in the bank vole suggest a different evolutionary history from other rodents

S Marková[1], JB Searle[2] and P Kotlík[1]

Gene duplication plays an important role in the origin of evolutionary novelties, but the mechanisms responsible for the retention and functional divergence of the duplicated copy are not fully understood. The α-globin genes provide an example of a gene family with different numbers of gene duplicates among rodents. Whereas *Rattus* and *Peromyscus* each have three adult α-globin genes (HBA-T1, HBA-T2 and HBA-T3), *Mus* has only two copies. High rates of amino acid evolution in the independently derived HBA-T3 genes of *Peromyscus* and *Rattus* have been attributed to positive selection. Using RACE PCR, reverse transcription-PCR (RT–PCR) and RNA-seq, we show that another rodent, the bank vole *Clethrionomys glareolus*, possesses three transcriptionally active α-globin genes. The bank vole HBA-T3 gene is distinguished from each HBA-T1 and HBA-T2 by 20 amino acids and is transcribed 23- and 4-fold lower than HBA-T1 and HBA-T2, respectively. Polypeptides corresponding to all three genes are detected by electrophoresis, demonstrating that the translated products of HBA-T3 are present in adult erythrocytes. Patterns of codon substitution and the presence of low-frequency null alleles suggest a postduplication relaxation of purifying selection on bank vole HBA-T3.
Heredity (2014) **113,** 64–73; doi:10.1038/hdy.2014.12; published online 5 March 2014

INTRODUCTION

Gene duplication has been in the focus of evolutionary biologists since the classic text of Ohno (1970) on *Evolution by Gene Duplication*. Although alternative mechanisms have been proposed for the origin of new genes (Kaessmann, 2010; Tautz and Domazet-Lošo, 2011; Ding *et al.*, 2012), it is clear that gene duplication is an important source of evolutionary novelty (Lynch and Conery, 2003; Zhang, 2003). Such duplication may involve the whole genome, chromosome segments, individual genes or only parts of genes (Betrán and Long, 2002; Kaessmann, 2010).

Much of the study of the evolutionary role of gene duplication has been focused on divergence among members of gene families (Robin *et al.*, 2000; Friedman and Austin, 2001; Storz *et al.*, 2011, 2013). However, the questions of how often such gene duplicates arise, by which mechanisms they are maintained and how frequently they evolve a new function remain insufficiently addressed (Zhang, 2003; Hahn, 2009; Innan and Kondrashov, 2010). A number of models for the evolution and retention of gene duplication have been proposed, but their relative importance is a matter of debate, and overall this is a poorly understood topic (summarised by, for example, Zhang, 2003; Innan and Kondrashov, 2010). Most models predict that amino acid substitutions are needed in one or both gene copies after the duplication for both genes to be stably maintained in the genome. Thus, according to the 'neofunctionalisation' model of Ohno (1970), the functional redundancy of the duplicate gene relieves it from negative purifying selection, and if not pseudogenised by nonsense mutations, it may accumulate neutral amino acid substitutions and acquire a new function that is further improved by positive selection (Innan and Kondrashov, 2010). Relaxation of purifying selection may also allow fixation of alternate mutations in both copies that can be damaging to the gene function so that neither copy is sufficient on its own to perform the original function, and both genes must be maintained by selection (the 'duplication-degeneration-complementation' model; Force *et al.*, 1999). Alternatively, different adaptive substitutions may be acquired by each paralogous copy, leading to partitioning of ancestral functions between the duplicates (Hughes, 1994). This 'escape of adaptive conflict' model posits that the original gene had multiple functions that could not be independently improved because of pleiotropic constraints, but could each be adopted and improved by one copy after the duplication (Hughes, 1994; Storz, 2009).

A common approach to assess selection pressure on protein-coding genes is to compare substitution rates at nonsynonymous and synonymous sites dN and dS, respectively (Lynch and Conery, 2000; Innan and Kondrashov, 2010). Recently, comparative genomic studies have used this approach to reveal genome-wide patterns for different organisms (Lynch and Conery, 2000; Colbourne *et al.*, 2011). For example, a comparison across several model eukaryotic species suggested that although most paralogous genes evolve under relaxed selective constraints shortly after the duplication (up to $dN/dS \approx 1$), most duplicates are inactivated before reaching a few percent divergence at synonymous sites, and those that remain functional

[1]Institute of Animal Physiology and Genetics, Academy of Sciences of the Czech Republic, Liběchov, Czech Republic and [2]Department of Ecology and Evolutionary Biology, Cornell University, Ithaca, NY, USA
Correspondence: Dr P Kotlík, Institute of Animal Physiology and Genetics, Academy of Sciences of the Czech Republic, Rumburská 89, Liběchov, Czech Republic.
E-mail: kotlik@iapg.cas.cz
Received 6 July 2013; revised 6 December 2013; accepted 10 December 2013; published online 5 March 2014

then evolve under strong purifying selection ($dN/dS \ll 1$) (Lynch and Conery, 2000). Such genomic studies thus have the potential to reveal overall trends. However, a detailed molecular characterisation of individual gene families remains necessary to uncover the mechanisms behind the evolution by gene duplication, including in nonmodel organisms.

The mammalian α-globin gene family represents one of the best-characterised set of genes originating by gene duplication (Tufarelli *et al.*, 2004; Hoffmann and Storz, 2007; Hoffmann *et al.*, 2008, 2010). The high rate of gene duplication and loss in this family is likely mediated by unequal crossing-over and inactivation (Hoffmann *et al.*, 2008). Many of the gene duplicates within a species therefore have no 1:1 orthologues in other species (see, for example, Campos *et al.*, 2012), and often have nearly identical sequence as the result of concerted evolution by means of interparalogue gene conversion (Hoffmann *et al.*, 2008). However, the α-globin genes of two rodents, the Norway rat (*Rattus norvegicus*) and the deer mouse (*Peromyscus maniculatus*), are a striking exception to this general pattern. Each species possesses three copies of adult α-globin genes, where two of the paralogues (HBA-T1 and HBA-T2) have almost identical sequences because of concerted evolution, whereas the third paralogue (HBA-T3) shows differences at multiple amino acid sites, predicted to produce changes in oxygen-binding affinity (for details see Hoffmann *et al.*, 2008; Storz *et al.*, 2008). The HBA-T3 paralogue was inferred to originate independently in *Rattus* and in *Peromyscus*, and in each species the high rate of protein evolution has been attributed to positive directional selection rather than relaxed purifying selection (Storz *et al.*, 2008). The haemoglobin differentiation in each of these species was thus considered to be driven by selection favouring a functional division of labour among haemoglobin isoforms that incorporate products of the different α-globin paralogues (Storz *et al.*, 2008).

It is interesting that in the house mouse (*Mus musculus*), the organisation of the α-globin gene cluster parallels that in *Rattus* and *Peromyscus*, but in mouse the third α-globin gene is a pseudogene that has been translocated to another chromosome (Tufarelli *et al.*, 2004). There are thus only two functional α-globin genes in mouse (HBA-T1 and HBA-T2). Therefore, the two murids (*Mus* and *Rattus*) appeared to differ by the number of functional HBA copies, with the situation in *Rattus* being more similar to the cricetid *Peromyscus*.

Intriguingly, a recent study of Storz *et al.* (2010) assessing the transcriptional activity of the *Peromyscus* HBA genes by using reverse transcription-PCR (RT–PCR) did not detect transcripts matching the HBA-T3 gene in the definitive erythrocytes of adult deer mice, suggesting the gene might be transcriptionally inactive although it has a complete open reading frame (Storz *et al.*, 2010).

Because of variation in the number of the functional copies in different murid species, the independent origin of functionally district triplicates in murid and cricetid species and the intriguing transcriptional pattern in *Peromyscus*, it is of interest to study the α-globin genes in other rodents. Here, we isolate and study the α-globin genes in a Eurasian cricetid species, the bank vole *Clethrionomys glareolus* (also known as *Myodes glareolus*; see Tesakov *et al.*, 2010). Specifically, we used RNA, DNA and protein analyses to: (1) isolate the bank vole α-globin genes by rapid amplification of cDNA ends (RACE) and sequence each paralogue in a large sample of *C. glareolus* and in representatives of other *Clethrionomys* species; (2) determine the transcriptional activity of each gene in adult bank voles using RT–PCR; (3) measure their relative levels of expression by RNA-seq; and (4) test whether the observed amino acid differences among α-globin genes in the bank vole are attributable to positive selection similar to HBA-T3 gene of *Rattus* and *Peromyscus*.

MATERIALS AND METHODS

Samples and DNA and RNA extraction

Altogether, 145 individuals of *C. glareolus* were collected from 12 sites in Britain, with an average of 12 specimens per site. In addition, one specimen of bank vole from Calabria, Italy (a likely sibling species to *C. glareolus*; Colangelo *et al.*, 2012), and one grey red-backed vole *C. rufocanus* from Norway were sampled. Blood was obtained by cardiac puncture from killed bank voles and the heparinised cells were collected by centrifugation (3000 r.p.m. in a microfuge, that is, $\sim 630 \times g$, 10 min), washed with phosphate-buffered saline, resuspended in water and frozen in liquid nitrogen. Samples of bone marrow and spleen (an erythropoietic organ in rodents; Brodsky *et al.*, 1966; Cheng *et al.*, 1974) were stored in RNAlater (Qiagen, Valencia, CA, USA) or 96% ethanol for DNA analyses. The genomic DNA was extracted using the DNeasy Tissue Kit (Qiagen). Total RNA was extracted from the samples stored in RNAlater with RNeasy Mini Kit (Qiagen), followed by DNase treatment using TURBO DNA-free Kit (Ambion, Austin, TX, USA) and a clean-up using the RNeasy Mini Kit.

Cloning and sequencing of α-globin genes

To amplify the α-globin genes from genomic DNA in the bank vole, we used published sets of PCR primers originally designed to amplify selectively the individual paralogous α-globin genes in *Mus* (Storz *et al.*, 2007b) and in *Peromyscus* (Storz *et al.*, 2007a). However, only the primer pair designed to selectively amplify the HBA-T2 gene in *Peromyscus* (D-1518 and D-2387R; Storz *et al.*, 2007a) yielded a specific PCR product in the bank vole. Therefore, to isolate any other α-globin genes from this species, we cloned and sequenced bank vole globin complementary DNA (cDNA). Total RNA was reverse transcribed using the High Capacity cDNA Reverse Transcription Kit (Applied Biosystems, Foster City, MA, USA). We first obtained partial bank vole cDNA sequences using a pair of primers matching highly conserved motifs in exons of mammalian α-globin genes (Yingzhong *et al.*, 2007). We then designed new primers in regions of the partial bank vole sequences obtained that did not show divergence (that is, overlapped multiple peaks in chromatograms) between paralogous genes co-amplified with the conserved primers. The primers were used for RACE PCR with the SMARTer RACE cDNA Amplification Kit and Advantage 2 PCR Kit (Clontech, Mountain View, CA, USA). This strategy ensured that transcripts of all transcriptionally active paralogous copies of bank vole α-globin genes were amplified. The RACE products were cloned with the Qiagen PCR Cloning plus Kit and sequenced from 55 to 103 clones per vole, and full-length cDNA sequence was assembled for each gene on the basis of overlaps in the 5′- and 3′-RACE sequences. From these sequences we designed gene-specific primers to selectively amplify and sequence each of the bank vole α-globin genes from genomic DNA. The primers were selected to match sequences of the 5′ and 3′untranslated regions (UTRs) that allowed each gene to be amplified and sequenced from initiation to termination codon. Two new forward primers (5′-ACACACTTCTGATTCTGAGA-3′ and 5′-CGGACTCAGGAAGGAATTCAT-3′) were used in combinations with the primer D-2387R (Storz *et al.*, 2007a) to amplify the HBA-T1 and HBA-T3 genes, respectively. Each PCR (total volume of 25 ml) contained 2.5 units of Easy-A high-fidelity PCR cloning enzyme (Agilent Technologies, La Jolla, CA, USA), 2.5 μl of 10 × Easy-A reaction buffer, 0.2 mM dNTP and 0.3 mM primers. The PCR amplification consisted of an initial denaturing at 94 °C for 2 min followed by 33 cycles of denaturing at 94 °C for 30 s, annealing at 58 °C for 40 s and extension at 72 °C for 60 s, with a final extension period of 10 min at 72 °C. All genotypes containing multiple heterozygous sites were resolved into haplotypes by cloning and sequencing of four to eight clones.

Transcriptional activity assessed by RT–PCR

We performed RT–PCR and PCR assays followed by agarose gel electrophoresis to verify the amplification of HBA-T3 cDNA and genomic DNA using the paralogue-specific primers. HBA-T1 cDNA served as a RT–PCR control. The PCR products were run on a 1.8% starch gel and the electrophoresis was carried out at 80 V for 3 h.

Separation of α-globin polypeptides

To assess the translational activity of the bank vole α-globin genes, we precipitated the globin chains with hydrochloric acid–acetone solution (see, for

example, Ferrand, 1989) and separated them by cellulose acetate membrane electrophoresis with a tris–borate–EDTA buffer (pH 8.9) in 8 M urea (Duffy *et al.*, 1976). The electrophoresis was carried out at 450 V for 105 min at 4 °C.

RNA-seq and expression analysis

Library preparation and sequencing were performed with standard Illumina (San Diego, CA, USA) protocols on the Illumina HiSeq 2000. Briefly, after poly(A) enrichment and fragmentation, the RNA was size selected to 250–400 bp, reverse transcribed into cDNA, end repaired and PCR enriched. The resulting libraries were sequenced using the 100-bp paired-end module. RNA-seq data from each vole were mapped to transcript references matching the sequences of the alleles at each α-globin gene of that particular vole using the CLC Genomics Workbench, version 6.0.1 (CLC bio A/S, Aarhus, Denmark). No mismatches were thus allowed that permitted RNA-seq data of each vole to be unambiguously assigned to a specific paralogue and allele. The number of mapped reads for each gene was determined and expression levels were calculated in units of reads per kb per million mapped reads (RPKM). A one-sided binomial test was used to test for differential abundance of transcripts produced by each allele in heterozygotes (allelic imbalance (AI)) following Fontanillas *et al.* (2010).

Gene conversion analysis

Signatures of gene conversion between the paralogous loci were assessed with the method of Betrán *et al.* (1997) that estimates the probability of each site being informative of a conversion event, and using the program GENECONV (Sawyer, 1999) that performs a permutation test to estimate the probability of observing unusually long stretches of concordant sites by chance (Sawyer, 1989).

Phylogenetic analysis

The phylogenetic relationships were estimated using the maximum-likelihood (ML) criterion and the combination of the NNI (nearest neighbour interchanges) and SPR (subtree pruning and regrafting) algorithms as implemented in GARLI, version 2.0 (Zwickl, 2006), with character partitions according to codon positions. Multiple GARLI runs were performed to ensure convergence on the same topology, each consisting of 10 replicates. Bayesian phylogenetic analysis was performed with MRBAYES, version 3.2.1 (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003), with character partitions according to codon positions. Two independent analyses were run to ensure convergence, each with four Markov chains under the Metropolis coupling. The length of each analysis was one million generations, with trees sampled every 100 generations and a burn-in of 250 trees. Neighbour-joining (NJ) trees were constructed from HKY + G corrected distance with PAUP*, version 4.0b10 (Swofford, 2003). The HKY-G model was determined to be the single best-fit model according to the Akaike information criterion by jModeltest, version 2.1.4. (Darriba *et al.*, 2012), and as the appropriate model for the first codon positions, and GTR + G model for the second and third codon positions. Branch support for the ML and NJ trees was assessed based on 1000 bootstrap resamplings.

In addition to the bank vole sequences, we also included published sequences of α-globin genes from the species analysed by Storz *et al.* (2008), that is, *P. maniculatus*, *R. norvegicus* and *M. musculus*. As in the study of Storz *et al.* (2008), a sequence of the single HBA gene of the guinea pig (*Cavia porcellus*) was used as outgroup (Fabre *et al.*, 2012).

Codon analysis of selection pressure

To evaluate the variation in selection regimes among different partitions of the rodent α-globin gene tree, we inferred ML estimates of the ω (dN/dS) ratio using the CODELM program of PAML, version 4.7 (Yang, 1997, 2007). We adopted a similar testing strategy to that of Storz *et al.* (2008). We first calculated a set of nested branch models, which allowed ω to vary among branches, and compared them through a likelihood ratio test. We first compared a single ω ratio model with a two-ratio model allowing a different ω for the branches of the HBA-T3 genes of all three species. To test whether the bank vole HBA-T3 gene experienced different selection regime from HBA-T3 genes of *Rattus* and *Peromyscus*, we compared the two-ratio model with a three-ratio model allowing a different ω for the HBA-T3 branch of *Clethrionomys* than for the HBA-T3 branches of *Rattus* and *Peromyscus*. In the third likelihood ratio test, we compared the three-ratio model with a four-ratio model with a different ω assigned to the HBA-T3 branch of each of the three species to test whether the HBA-T3 branch may have been under different selection regime in each species. We then applied a branch-site test of positive selection (test 2; Yang *et al.*, 2005) to evaluate whether positive selection affected individual codon sites along the HBA-T3 branch of *Clethrionomys*, labelled here as the foreground branch. A likelihood ratio test was used to compare the model A that assumed four site classes (class 0 of codons under stringent functional constraints with $0<\omega_0<1$; class 1 of unconstrained codons with $\omega_1=1$; and classes 2a and 2b of codons conserved or neutral on the background branches, but that could be under positive selection with ω_2 on the foreground branch), with the null model fixing ω_2 at 1. In addition, we compared clade model C (Bielawski and Yang, 2004) having four free ω parameters (ω_2, ω_3, ω_4 and ω_5) against the M1a (Nearly Neutral) model considering only two types of sites for the entire tree with $0<\omega_0<1$ and $\omega_1=1$ (Nielsen and Yang, 1998; Yang *et al.*, 2005) that allowed us to specify the HBA-T3 branches of *Rattus*, *Peromyscus* and *Clethrionomys* as independent foreground branches.

As an alternative approach, we used the HyPhy package, accessible through the Datamonkey web (http://www.datamonkey.org./) interface (Kosakovsky Pond *et al.*, 2005). Unlike PAML, the genetic algorithm branch analysis does requires *a priori* specification of a branch suspected to show an independent pattern of molecular evolution, but aims to find the best-fitting branch model for the data (Kosakovsky Pond and Frost, 2005). The branches are sorted into a number of rate classes and a genetic algorithm is used to find the best-fitting branch model among all those that allocate each tree branch to one of the rate classes, with a separate ω estimated for each class. The goodness of fit of each model is determined by the Akaike information criterion and the model-averaged probability is calculated for each branch of $\omega>1$ (Kosakovsky Pond and Frost, 2005).

Phylogenetically independent estimates were obtained by comparing the levels of divergence and polymorphism at nonsynonymous and synonymous sites. Pairwise divergence among the genes was estimated using PAML. The polymorphism level was measured by the average number of pairwise differences (π) among all 145 bank voles and was calculated with DnaSP, version 5.10.01 (Librado and Rozas, 2009).

Structural modelling and conservation analysis

A homology-based model of bank vole haemoglobin was predicted and built by SWISS-MODEL (Arnold *et al.*, 2006) using the experimentally determined structure of mouse haemoglobin (3HRW) as the template. Visualisations and renderings were done with PyMOL Molecular Graphics System, version 1.5.0.3. (Schrödinger, LLC, Portland, OR, USA).

To evaluate the degree of conservation of α-globin sites across mammals, we estimated protein conservation scores for each site with the method of Ashkenazy *et al.* (2010) accessible through the web server ConSurf (http://consurf.tau.ac.il). We compiled an alignment of 186 published α-globin protein sequences from 128 mammalian species and used ConSurf to compute the site-specific conservation scores applying a Bayesian algorithm (Ashkenazy *et al.*, 2010). The estimated conservation scores were then divided into a discrete scale of nine colour grades for visualisation (from the least conserved sites corresponding to grade 1 to the most highly conserved sites corresponding to grade 9).

We used the CLC Genomics Workbench to predict the net charge differences between the globin polypeptides under the pH conditions used for the electrophoresis.

RESULTS

Cloning and characterisation of α-globin genes in bank vole

The sequences of cloned RACE products unambiguously assembled into full-length cDNA sequences of three genes, each with an intact open reading frame of 141 amino acids, starting with an AUG Met codon and terminating with a UAA stop codon, and flanked by a 37-bp (for HBA-T1 and HBA-T2) to 38-bp (for HBA-T3) UTR

upstream and a 91-bp UTR downstream. The 3′UTR of each sequenced transcript showed a polyadenylation signal site (AAUAAA) followed by a poly(A) tail. BLAST (Basic Local Alignment Search Tool; Altschul *et al*., 1990) matches for all three genes confirmed they are α-globin genes. The cDNA sequence of one of the bank vole genes matched the sequence of the gene amplified from bank vole genomic DNA with the primers D-1518 and D-2387R (Storz *et al*., 2007a), indicating the orthology of this bank vole gene with *Peromyscus* HBA-T2. The cDNA sequence of the second bank vole gene showed very high sequence similarity (97%) to the bank vole HBA-T2 gene and we thus tentatively refer to it as HBA-T1 (Figure 1). The third gene was distinguished from the other two genes by a high number of amino differences, and we refer to it as HBA-T3. The bank vole HBA-T3 has 19 unique amino acid differences compared with HBA-T1 and HBA-T2, plus it differs from each by another additional amino acid (Figure 1). Among the 93 5′-RACE PCR clones sequenced for 3 voles, 77 clones were derived from the HBA-T1 transcript, 12 from the HBA-T2 transcript and 4 from the HBA-T3 transcript. The sequences of PCR products obtained from the genomic DNA with paralogue-specific primers contained exon–intron splice junctions matching those found in the α-globin genes of other mammals, with three exons and two introns. The sequence of the coding region (exclusive of introns) of each gene matched the corresponding transcript sequence of the same vole. On the basis of these results, we concluded that the bank vole has three copies of transcriptionally active α-globin genes that we isolated and separately amplified with specific primers.

HBA-T3 alleles containing premature termination codons

The sequencing survey of 145 bank voles from Britain revealed two HBA-T3 alleles with truncated open reading frames because of the presence of a premature termination codon (PTC). Each allele was recovered from two individuals and each was present in a homozygous state as well as in a heterozygous state with an allele with an intact open reading frame. One of the alleles had a PTC in exon 2 (hereafter PTC1) and the second allele had a PTC in exon 3 (PTC2) (Figures 1 and 2a). PCR amplification from cDNA using a primer pair specific for HBA-T3 produced an amplicon of the expected size for the voles homozygous and heterozygous for the PTC2 allele, but only for the heterozygous individual in the case of the PTC1 allele (Figure 2b). No detectable amplicon was obtained for the vole homozygous for the PTC1 allele (Figure 2b). In contrast, PCR amplification from the same cDNA, but using a primer pair specific for HBA-T1, yielded the expected amplicon in all voles (Figure 2b).

Electrophoretic analysis (Figure 2c) showed that the HBA-T3 polypeptide was the most negatively charged subunit (predicted net charge of −5), followed by β-globin (−1), HBA-T1 (0) and HBA-T2 (+2) polypeptides. The high staining intensity of the HBA-T1 polypeptide, medium staining intensity of the HBA-T2 and low staining intensity of the HBA-T3 polypeptide (Figure 2c) correspond with the relative expression levels of the three genes as identified by RNA-seq (Figure 3a). Importantly, there was no detectable staining of the HBA-T3 polypeptide in the two voles homozygous for either the PTC1 or the PTC2 allele (Figure 2c, individuals B/B and C/C), although the products were clearly present in both heterozygous voles (individuals A/B and A/C).

Differential gene expression using RNA-seq

We collected RNA-seq data for three voles that were homozygous at all three α-globin genes. The data confirmed that all three HBA paralogues are transcriptionally active genes in the bank vole. The expression levels revealed large differences in transcript abundance among the three genes. The mean RPKM values were 1.9 million for HBA-T1, 325 000 for HBA-T2 and 84 000 for the HBA-T3 transcript

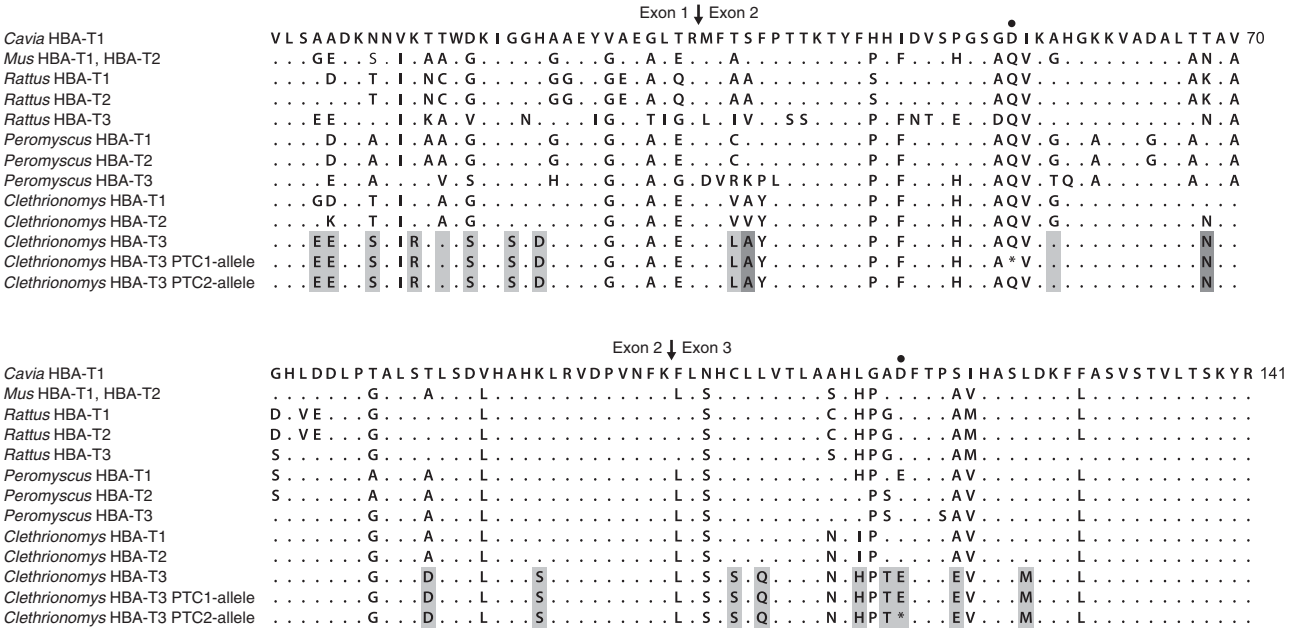

Figure 1 Deduced amino acid sequences of bank vole (*Clethrionomys*) α-globins aligned with those from other rodents. Dots represent amino acids identical to those at corresponding positions in *Cavia* sequence. Residues distinguishing bank vole HBA-T3 from bank vole HBA-T1 and HBA-T2 are shaded in grey, with residues distinguishing HBA-T3 from only one of the two genes in dark grey. Arrows indicate exon–exon junctions. The position of single nonsense mutations in two bank vole HBA-T3 alleles that generate in-frame PTCs are marked by a dot. The transcript of the PTC1 allele is most likely subject to the nonsense-mediated mRNA decay, whereas the PTC2 allele appears capable of producing a truncated α-globin; the read-through sequences are provided for comparison.

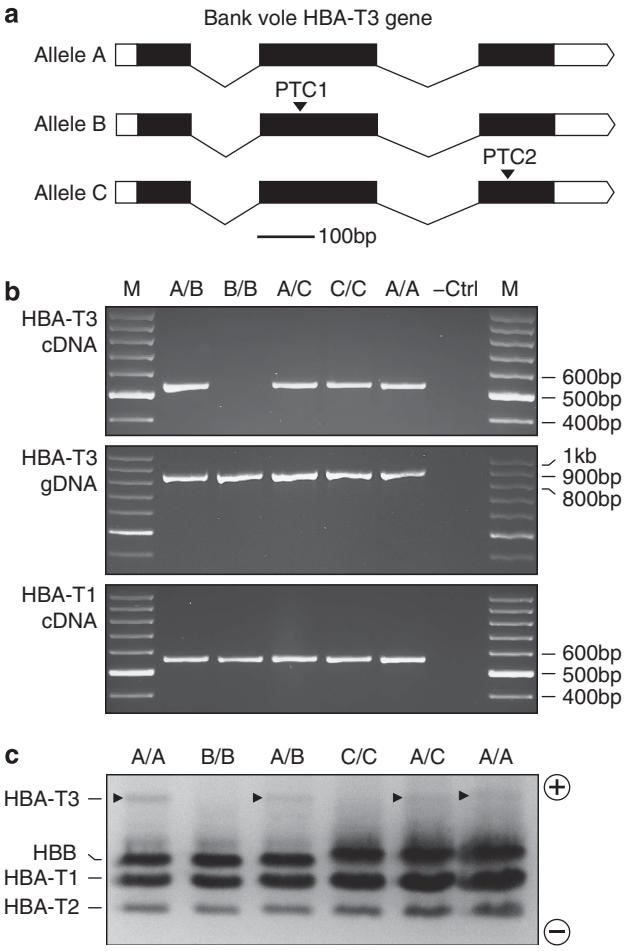

Figure 2 RT–PCR and protein expression analysis of the bank vole HBA-T3 globin gene. (**a**) Genomic structure of the gene. Alleles containing PTCs in exons 2 and 3, referred to as, respectively, the B allele and C allele, are aligned with a non-PTC allele referred to as the A allele. (**b**) RT–PCR and PCR assays followed by gel electrophoresis were performed to assess the amplification of HBA-T3 cDNA and genomic DNA (gDNA) using paralogue-specific primers spanning the entire coding region. Data represent HBA-T3 cDNA amplification from mRNA of all genotypes except of the PTC1 homozygote. The HBA-T1 cDNA served as a RT–PCR control and was amplified in all individuals. (**c**) Urea cellulose acetate electrophoresis of globin polypeptides demonstrating the absence of a HBA-T3 translation product in the homozygotes for both PTC-containing alleles, whereas the heterozygotes and the individuals with non-PTC genotypes show a detectable (although weakly staining) band (arrows).

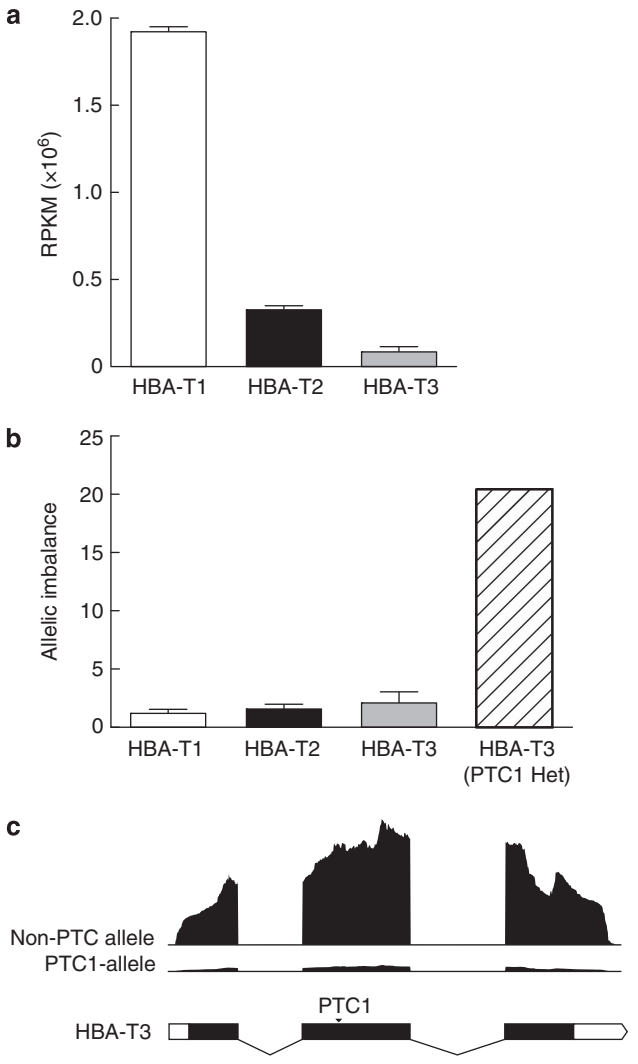

Figure 3 RNA-seq expression profiling of the bank vole α-globin genes. (**a**) The mean RPKM of each HBA paralogue with error bars representing s.e.m. (**b**) Fold difference in transcript abundance between two alleles for each HBA paralogue (that is, AI). In the first three columns, the error bar represents s.e.m. The rightmost column (hatched) shows the result for an individual heterozygous (Het) for the allele containing a premature stop codon in exon 2 (PTC1). (**c**) RNA-seq depth-of-coverage profiles of an allele with an uninterrupted open reading frame and of the PTC1-containing allele from the heterozygote.

reference, indicating that the HBA-T1 gene is expressed approximately sixfold more than the HBA-T2 gene and 23-fold more than the HBA-T3 gene, and that the HBA-T2 gene is expressed approximately fourfold more highly than the HBA-T3 gene (Figure 3a).

AI in PTC heterozygote

Quantification of allele-specific expression revealed detectable AI in all the three α-globin genes (Figure 3b). Along with the vole heterozygous for the PTC1-containing allele at HBA-T3, we tested the AI of HBA-T1 and HBA-T3 in three voles heterozygous at these genes and of HBA-T2 in four voles. Significant AI of all three genes was detected in all voles (binomial test: $P<0.001$), with a mean of 1.19 for HBA-T1, 1.57 for HBA-T2 and 2.08 for HBA-T3 (including only voles with non-PTC alleles; Figure 3b). The transcript abundance of the non-PTC allele in the vole heterozygous for the PTC1 allele at HBA-T3 was 37 534 RPKM, whereas that of the PTC1 allele was only 1846 RPKM. The vole heterozygous for the PTC1 allele at HBA-T3 therefore showed a much larger AI of HBA-T3 (20.33) than the other voles had at any of the three genes (Figures 3b and c).

Gene conversion

We detected nine gene conversion events between the bank vole HBA-T1 and HBA-T2 using the method of Betrán *et al.* (1997), six in four HBA-T1 alleles (three in the same allele) and three in HBA-T2 (each in a different allele). The conversion tracts spanned both exons and introns, with the exception of exon 1 where no gene conversion tracts were detected. The median conversion tract length was 13 bp in HBA-T1 and 134 bp in HBA-T2, but it varied broadly in both genes, from 2 to 48 bp in HBA-T1 and from 20 to 134 bp in HBA-T2.

The GENECONV analysis identified four additional conversion tracts in HBA-T2, each in a different allele, three of 614 bp and one of 696 bp. The tracts spanned exon 2, intron 2 and exon 3 of the gene. We however found only one conversion involving HBA-T3, a tract of 21 bp in exon 2 of HBA-T2 identified by the method of Betrán *et al.* (1997). Although conversion tracts in some alleles might have a common origin, our results suggest that reciprocal gene conversion has occurred frequently between bank vole HBA-T1 and HBA-T2 genes, whereas HBA-T3 appears to be more rarely involved and mostly as the donor gene.

Phylogenetic relationships of rodent α-globin genes

Combining the sequences of the coding region of each bank vole α-globin gene with those of other rodents (Storz *et al.*, 2008) resulted in a data set containing 169 variable sites, 97 of which were phylogenetically informative. The ML tree estimate showing the ML and NJ bootstrap values and the Bayesian posterior probabilities is presented in Figure 4a. Phylogenetic reconstructions obtained with the NJ and Bayesian approaches yielded similar topologies to the ML tree, and although the branching order between species was not resolved with high confidence, all analyses consistently placed the α-globin sequences from the same species in a group exclusive of the sequences from other species (Figure 4a), corroborating the results of Storz *et al.* (2008), who attributed this finding to a history of concerted evolution between HBA-T1 and HBA-T2 within each species. Additional phylogenetic analyses including other *Clethrionomys* voles placed HBA-T3 genes of the British bank vole, the Calabrian bank vole and the grey red-backed vole into one subclade (Supplementary Figure S1).

Variation in selection pressure on HBA-T3 among rodents

The likelihood ratio test rejected the one-ratio branch model assuming the same ω for all branches in the α-globin phylogeny ($P<0.001$), and indicated that the three-ratio model, allowing a different ω for the HBA-T3 branch of *Clethrionomys* than for the HBA-T3 branches of *Peromyscus* and *Rattus*, provided a significantly better fit than the two-ratio model assuming the same ω ratio for all three HBA-T3 branches ($P<0.01$; Table 1). The estimate $\omega=0.33$ was obtained for the bank vole HBA-T3 branch, $\omega=1.25$ for the HBA-T3 branches of *Peromyscus* and *Rattus* and $\omega=0.20$ for the other branches in the three-ratio model. The four-ratio model, allowing a different ω for the HBA-T3 branch of each *Clethrionomys*, *Peromyscus* and *Rattus*, did not fit significantly better than the three-ratio model ($P>0.1$; Table 1). These results hold under the Bonferroni adjustment of significance level for multiple comparisons ($0.05/3=0.017$). The branch-site test of positive selection (Yang *et al.*, 2005; Zhang *et al.*, 2005), aimed at testing whether some of the codon sites along the bank vole HBA-T3 branch might evolve by positive selection (that is, at $\omega>1$), was nonsignificant ($P>0.5$; Table 1). The clade model C (Bielawski and Yang, 2004), which allowed a subset of codon sites along each of the three HBA-T3 branches to evolve under positive selection, with different ω estimated for each branch, was a significantly better fit than the Nearly Neutral Model M1a ($P<0.001$; Table 1). However, only in the HBA-T3 branches of *Peromyscus* and *Rattus* was there a class of sites with $\omega>1$ (1.54 and 2.89, respectively), whereas for the bank vole HBA-T3 branch, $\omega=0.41$ was estimated for the same site class (Table 1).

The model selected by the genetic algorithm branch analysis sorted the branches of the α-globin tree into four rate classes, with the ω estimated for each class ranging from $\omega=0.14$ to $\omega\gg1$ (Figure 4b). Consistent with the PAML results, the HBA-T3 branches of *Peromyscus* and *Rattus* were allocated to a rate class with $\omega=1.37$, with the model-averaged probability of $\omega>1$ exceeding 0.99 for these two branches, whereas the bank vole HBA-T3 branch was allocated to a rate class with $\omega=0.39$ and had zero probability of $\omega>1$ (Figure 4b).

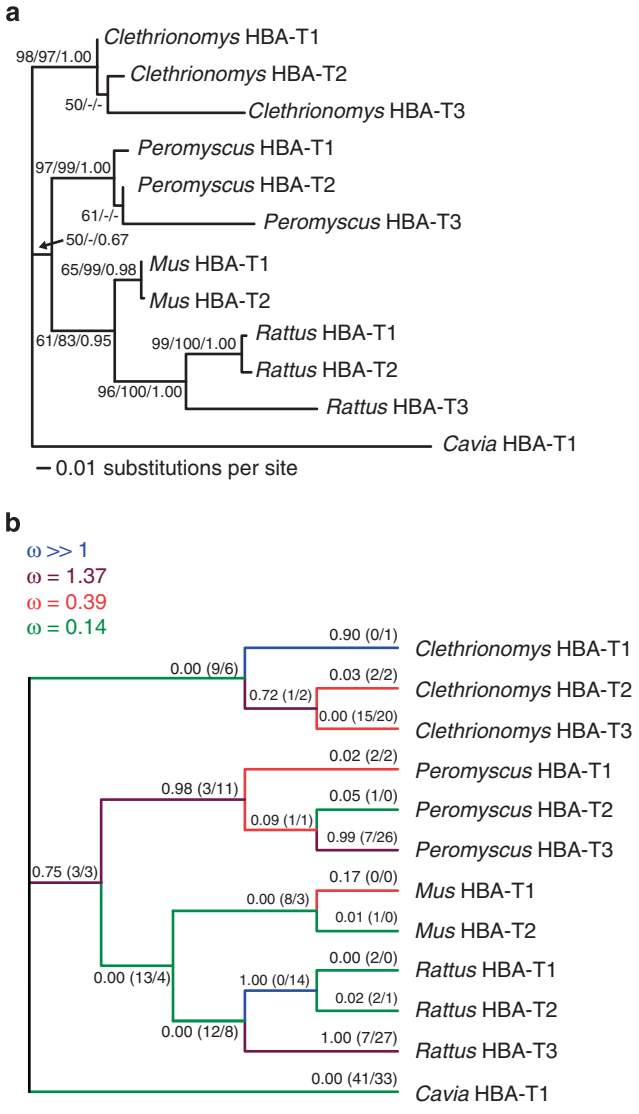

Figure 4 Phylogenetic relationships among rodent α-globin genes. (**a**) ML tree inferred from coding sequences, with statistical support for tree bipartitions expressed as percentage bootstrap values from ML/NJ analyses and as the Bayesian posterior probabilities (values <50% not shown). (**b**) Variation in selection pressure across branches as inferred by the genetic algorithm branch analysis. Branch labels represent model-averaged probabilities of $\omega>1$ with the inferred number of synonymous and nonsynonymous substitutions in parentheses (NS/NN).

Concordant results, suggesting a reduced rate of evolution at functional sites in the bank vole HBA-T3, were obtained by phylogeny-independent estimates of the ratio of the nonsynonymous/synonymous divergence of HBA-T3 from HBA-T1 (0.44) and HBA-T2 (0.40), and when considering the levels of HBA-T3 polymorphism segregating at nonsynonymous ($\pi_N=0.0016$) versus synonymous sites ($\pi_S=0.0046$) among all 145 bank voles that yielded $\omega_\pi=0.35$.

Table 1 Codon analysis of selection on HBA-T3 genes

Model	*Parameter estimates*	*No. of parameters*	*l*	*2Δl*	P-*value*
Branch tests					
One-ratio model	$\omega_0 = 0.31$	23	−1924.23		
Two-ratio model	$\omega_0 = 0.20$, $\omega_{T3\ Clethr} = \omega_{T3\ Pero} = \omega_{T3\ Rat} = 0.72$	24	−1913.05	22.37	0.000
Three-ratio model	$\omega_0 = 0.20$, $\omega_{T3\ Clethr} = 0.33$, $\omega_{T3\ Pero} = \omega_{T3\ Rat} = 1.25$	25	−1909.26	7.58	0.006
Four-ratio model	$\omega_0 = 0.20$, $\omega_{T3\ Clethr} = 0.33$, $\omega_{T3\ Pero} = 1.01$, $\omega_{T3\ Rat} = 1.63$	26	−1909.01	0.50	0.481
Branch-site tests					
Model A0	$P_0 = 0.60$, $P_1 = 0.24$, $P_{2a} = 0.11$, $P_{2b} = 0.05$ Background: $\omega_0 = 0.05$, $\omega_1 = 1.00$, $\omega_{2a} = 0.05$, $\omega_{2b} = 1.00$ HBA-T3$_{Clethr}$: $\omega_0 = 0.05$, $\omega_1 = 1.00$, $\omega_{2a} = 1.00$, $\omega_{2b} = 1.00$	25	−1860.24		
Model A	$P_0 = 0.66$, $P_1 = 0.26$, $P_{2a} = 0.06$, $P_{2b} = 0.02$ Background: $\omega_0 = 0.06$, $\omega_1 = 1.00$, $\omega_{2a} = 0.06$, $\omega_{2b} = 1.00$ HBA-T3$_{Clethr}$: $\omega_0 = 0.06$, $\omega_1 = 1.00$, $\omega_{2a} = 2.36$, $\omega_{2b} = 2.36$	26	−1860.09	0.30	0.583
Neutral model M1a	$P_0 = 0.71$, $P_1 = 0.29$ $\omega_0 = 0.07$, $\omega_1 = 1.00$	24	−1862.40		
Model C	$P_0 = 0.56$, $P_1 = 0.20$, $P_2 = 0.25$ Background: $\omega_0 = 0.02$, $\omega_1 = 1.00$, $\omega_2 = 0.16$ HBA-T3$_{Clethr}$: $\omega_0 = 0.02$, $\omega_1 = 1.00$, $\omega_5 = 0.41$ HBA-T3$_{Pero}$: $\omega_0 = 0.02$, $\omega_1 = 1.00$, $\omega_4 = 2.89$ HBA-T3$_{Rat}$: $\omega_0 = 0.02$, $\omega_1 = 1.00$, $\omega_3 = 1.54$	29	−1849.67	25.46	0.000

The bank vole HBA-T1 and HBA-T2 genes differ from each other at four amino acid sites, whereas each of these genes differs from the bank vole HBA-T3 by 20 amino acid substitutions (Figure 1). Of these substitutions, 19 were inferred to have occurred along the HBA-T3 branch, but the Bayes Empirical Bayes analysis (Yang *et al.*, 2005) suggested that none of the substitutions is attributable to positive selection using the 95% posterior probability cutoff, and none occurs at a highly conserved site (conservation score of 9) as determined by the ConSurf method (Ashkenazy *et al.*, 2010). Instead, the majority of the substitutions distinguishing HBA-T3 in the bank vole (Figure 1) are at sites that are among the most variable in mammals (conservation score of 1; Figure 5).

Therefore, in summary, the HBA-T3 gene of the bank vole shows many amino acid differences with the two paralogous genes, similar to the situation in *Peromyscus* and *Rattus*, but only in these latter two species are the differences attributable to positive selection acting on HBA-T3, whereas no evidence of positive selection was found for HBA-T3 of the bank vole that, instead, appears to be under negative selection.

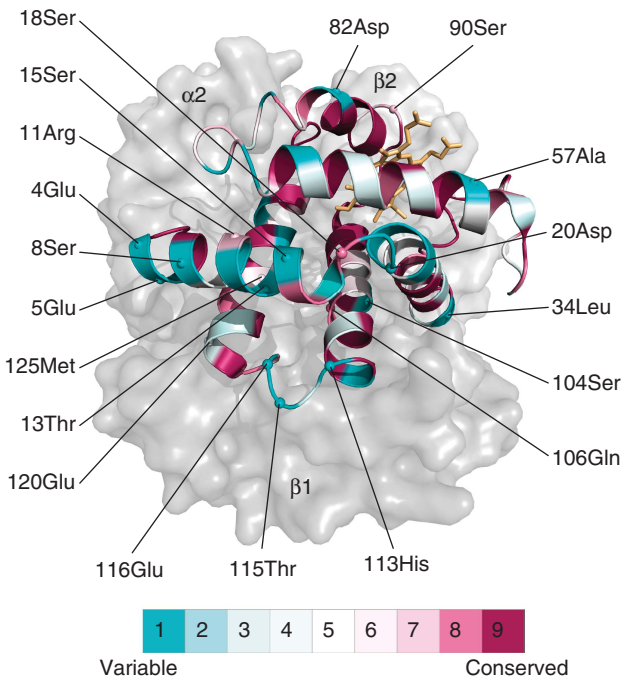

Figure 5 Homology-based structural model of the bank vole α-globin chain. The locations of 19 amino acids distinguishing the product of HBA-T3 from the products of HBA-T1 and HBA-T2 paralogues are shown. Residues are coloured according to their ConSurf evolutionary conservation score.

DISCUSSION

We have shown that the bank vole possesses three α-globin genes that are all transcriptionally active and produce functional α-globin polypeptides. Even though measures of transcript abundance may not necessarily be accurate predictors of protein abundance, the contribution of the individual paralogues to the synthesis of α-globin is clearly not equal (Figure 2c), with HBA-T1 showing an approximately sixfold and 23-fold higher transcript abundance than HBA-T2 and HBA-T3, respectively. Thus, the bank vole possesses a triplicate HBA-T3 gene with a high number of amino acid differences compared with HBA-T1 and HBA-T2, in which it is similar to *Peromyscus* and *Rattus*. However, everything suggests that in the bank vole the third paralogue is a functional gene whose translated products are present in adult erythrocytes. In *Peromyscus* and *Rattus*, the accelerated rates of amino acid substitution at the HBA-T3 gene

appear attributable to a positive directional selection, with some of the amino acid substitutions predicted to have important functional consequences in terms of oxygen-binding affinity (Storz *et al.*, 2008). It is therefore intriguing that no haemoglobin isoforms incorporating the products of HBA-T3 were detected in the definitive erythrocytes of adult *Peromyscus* (Storz *et al.*, 2010). The application of codon-substitution models revealed that the divergence of the bank vole HBA-T3 gene has been governed by purifying selection. We found no evidence of positive selection acting on the HBA-T3 in the bank vole. The ratio $\omega<1$ was consistently estimated for HBA-T3 in the bank vole and the majority of the amino acid differences are at sites predicted to be least conserved among mammals.

The dN/dS ratio estimated for the bank vole HBA-T3 branch was ~1.5-fold higher compared with the background value ($\omega=0.33$ vs $\omega=0.20$; Table 1), which is suggestive of a relaxation of purifying selection at HBA-T3. Based on the similarities of the flanking sequences and sharing of repetitive elements, the HBA-T3 genes appears to have originated by independent duplications in *Peromyscus* and *Rattus* (Storz *et al.*, 2008). Our RT–PCR strategy of isolating the bank vole HBA genes ensured that we did not miss any expressed gene, but it did not permit us to assess the orthologous relationships in a way similar to Storz *et al.* (2008), as we have no sequence information for the HBA-T3 flanking regions beyond the UTRs. Given the ubiquity of gene duplication and loss in the mammalian α-globin gene family (Hoffmann *et al.*, 2008), it is plausible that bank vole HBA-T3 represents an independent duplication from those in *Peromyscus* and *Rattus*. The fact that the Calabrian bank vole (a likely sibling species to *C. glareolus*; Colangelo *et al.*, 2012) and the grey red-backed vole (*C. rufocanus*) both share an orthologous HBA-T3 gene with the bank vole (Supplementary Figure S1) shows its origin predates the split between the ancestors of the bank vole and grey red-backed vole. The bank vole and grey red-backed vole belong to different major subdivisions in the genus (Cook *et al.*, 2004), and the duplication thus might be shared by all *Clethrionomys* species.

The finding of two PTC-containing HBA-T3 alleles in the bank vole, including in the homozygous state, is consistent with the idea that the gene is under somewhat relaxed functional constraints. Our RT–PCR experiments with primers spanning the entire coding region found evidence of an intact transcript from the allele with the PTC located in the third exon (Figure 2b). Undetectable RT–PCR amplification for the PTC1 allele (Figure 2b) and the 20-fold lower abundance of RNA-seq reads matching the PTC1 allele in the heterozygote suggest significantly lowered transcript abundance for the PTC1 allele. Because of the location of the PTC within the second of the three exons, the transcript of the PTC1 allele would likely be subject to the nonsense-mediated mRNA decay, a control mechanism in mammalian cells degrading defective mRNAs that would otherwise yield incomplete polypeptides (Baker and Parker, 2004). Only transcripts containing a termination codon more than ~50 nucleotides upstream of an exon–exon junction trigger nonsense-mediated mRNA decay (Maquat, 2004), which explains the amplification of the intact PTC2 allele transcript. Production of a truncated polypeptide by a physiologically important globin gene would undoubtedly be deleterious and result in a haemoglobin disorder (Thein, 2004), and hence there would be strong selection against the mutation concerned. For example, a PTC within the third exon of the human HBB gene does not trigger transcript degradation by nonsense-mediated mRNA decay because it is not followed by an exon–exon junction downstream (Thein, 2004; Maquat, 2005). The synthesis of a truncated and therefore nonfunctional β-globin causes a dominant form of thalassaemia, whereas PTCs located elsewhere than the last exon only result in a recessive haploinsufficiency (Thein, 2004; Maquat, 2005). The fact that in bank voles the PTC alleles were each found in a single population (each allele in a different population), and each in only two voles, points to a limited geographic and population spread of the PTC alleles. However, that both alleles were present in the heterozygous as well as homozygous states suggests that nonsense mutations in the bank vole HBA-T3 are not strongly deleterious. Therefore, the bank vole HBA-T3 appears functionally redundant relative to HBA-T1 and HBA-T2 to the extent that HBA-T3 inactivation does not have a strong negative physiological effect. Because it does not elicit nonsense-mediated mRNA decay, the PTC2 allele appears capable of producing a truncated α-globin, although at low abundance because of the low HBA-T3 expression level relative to HBA-T1 and HBA-T2.

Taken altogether, the results suggest that the bank vole HBA-T3 gene has been experiencing relaxed functional constraints. The higher ω value in HBA-T3 is consistent with a relaxation of purifying selection. However, ω is still significantly lower than 1, indicating that it is still subject to some degree of functional constraint. On the other hand, the strength of negative selection is not sufficient as to remove even silencing mutations and mutations resulting in nonfunctional protein, at a rate preventing their spread in the population. We therefore conclude that after the duplication giving rise to HBA-T3, negative selection has been relaxed in the new gene copy. The triplicate α-globin state is apparently fixed in the bank vole as we successfully amplified all these paralogues in each of the 145 bank voles and also in other *Clethrionomys*. However, the relaxation of negative selection on the HBA-T3 amino acid sequence suggests that the gene might not have reached the 'preservation phase', that is, the stage when the new gene performs a different function from the original gene, either because of a neofunctionalisation of the new gene or specialisation or subfunctionalisation of the two copies (Zhang, 2003; Innan and Kondrashov, 2010). Only after reaching this stage are both copies stably maintained by selection (Pegueroles *et al.*, 2013).

The functional significance of the bank vole HBA-T3 is currently unknown. The study of Wołk (1983) found three distinct types of haemoglobin in bank voles from Poland with different electrophoretic mobility and staining intensity. Two types were present in all individuals, presumably corresponding to molecules incorporating the products of HBA-T1 and HBA-T2, respectively (two bank vole HBB genes code for identical polypeptides; P Kotlík *et al.*, unpublished results). However, the third, the weakest staining and the most electronegative type, consistently appeared in voles of age >15 days, an important point in bank vole postnatal life when they develop open eyes, hearing, fur and the capacity for independent feeding (Wołk, 1983). The ontogenetic expression of the bank vole HBA-T3 remains to be established, but it appears plausible that its function might supplement HBA-T1 and HBA-T2, for example, when the transition to independent life poses new physiological demands. If HBA-T3 is mainly expressed during prenatal or early post-natal development of *Peromyscus*, then this could explain why a signature of positive selection is detected in spite of the fact that the product of HBA-T3 is not incorporated into functional haemoglobin tetramers in adult erythrocytes (Storz *et al.*, 2010).

Our study provides a glimpse at the critical stage of life of a gene duplicate before it either loses functional redundancy or is pseudogenised. It underscores the importance of the triplicated rodent α-globin genes as the model for the study of gene duplication, not only of functional structural divergence, but also of the evolutionary forces determining the fate of gene duplicates. The triplicate bank vole α-globin paralogue appears to be at a different 'life stage' from the

triplicate α-globin paralogues of *Rattus* and *Peromyscus*, and/or its functional divergence is of a different type from these species. Rather than having functionally significant structural changes, it may provide dosage supplementation during a physiologically demanding ontogenetic period. Therefore, in different rodent species the triplicate α-globin paralogues may be on different molecular and evolutionary trajectories.

DATA ARCHIVING

Nucleotide sequences for each gene identified in this study are available in the GenBank database (accession numbers: KF958827–KF958837).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

The work was supported by the Grant Agency of the Academy of Sciences of the Czech Republic (Grant Number IAA600450901), the Czech Science Foundation (Grant Number P506-11-1872) and the institutional support (RVO 67985904). Jana Kopecká, Vlastimil Šlechta and Antonín Stratil provided expert technical assistance. Gabriela Aguileta provided advice with branch tests and Sergei Kosakovsky Pond with HyPhy software. Gaetano Aloise and Giovanni Amori supported the fieldwork in Italy.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

Arnold K, Bordoli L, Kopp J, Schwede T (2006). The SWISS-MODEL Workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* **22**: 195–201.

Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N (2010). ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res* **38**: W529–W533.

Baker KE, Parker R (2004). Nonsense-mediated mRNA decay: terminating erroneous gene expression. *Curr Opin Cell Biol* **16**: 293.

Betrán E, Long M (2002). Expansion of genome coding regions by acquisition of new genes. *Genetica* **115**: 65–80.

Betrán E, Rozas J, Navarro A, Barbadilla A (1997). The estimation of the number and the length distribution of gene conversion tracts from population DNA sequence data. *Genetics* **146**: 89–99.

Bielawski JP, Yang Z (2004). A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution. *J Mol Evol* **59**: 121–132.

Brodsky I, Dennis LH, Kahn SB, Brady LW (1966). Normal mouse erythropoiesis. I. The role of the spleen in mouse erythropoiesis. *Cancer Res* **26**: 198–201.

Campos R, Storz JF, Ferrand N (2012). Copy number polymorphism in the α-globin gene cluster of European rabbit (*Oryctolagus cuniculus*). *Heredity* **108**: 531–536.

Cheng TC, Polmar SK, Kazazian HH (1974). Isolation and characterization of modified globin messenger ribonucleic acid from erythropoietic mouse spleen. *J Biol Chem* **249**: 1781–1788.

Colangelo P, Aloise G, Franchini P, Amori G (2012). Mitochondrial DNA reveals hidden diversity and an ancestral lineage of the bank vole in the Italian peninsula. *J Zool* **287**: 41–52.

Colbourne JK, Pfrender ME, Gilbert D, Thomas WK, Tucker A, Oakley TH *et al.* (2011). The ecoresponsive genome of *Daphnia pulex*. *Science* **331**: 555–561.

Cook JA, Runck AM, Conroy CJ (2004). Historical biogeography at the crossroads of the northern continents: molecular phylogenetics of red-backed voles (Rodentia: Arvicolinae). *Mol Phylogenet Evol* **30**: 767–777.

Darriba D, Taboada GL, Doallo R, Posada D (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods* **9**: 772.

Ding Y, Zhou Q, Wang W (2012). Origins of new genes and evolution of their novel functions. *Annu Rev Ecol Evol Syst* **43**: 345–363.

Duffy LK, Genaux CT, Stratton LP (1976). Amino acid differences between the α-chains from two hemoglobins of the yellow-cheeked vole (family Cricetidae). *Biochem Genet* **14**: 809–821.

Fabre P-H, Hautier L, Dimitrov D, Douzery EJP (2012). A glimpse on the pattern of rodent diversification: a phylogenetic approach. *BMC Evol Biol* **12**: 88.

Ferrand N (1989). Biochemical and genetic studies on rabbit hemoglobin. I. Electrophoretic polymorphism of the beta chain. *Biochem Genet* **27**: 673–678.

Fontanillas P, Landry CR, Wittkopp PJ, Russ C, Gruber JD, Nusbaum C *et al.* (2010). Key considerations for measuring allelic expression on a genomic scale using high-throughput sequencing. *Mol Ecol* **19**: 212–227.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**: 1531–1545.

Friedman R, Austin LH (2001). Pattern and timing of gene duplication in animal genomes. *Genome Res* **11**: 1842–1847.

Hahn MW (2009). Distinguishing among evolutionary models for the maintenance of gene duplicates. *J Hered* **100**: 605–617.

Hoffmann FG, Opazo JC, Storz JF (2008). Rapid rates of lineage specific gene duplication and deletion in the α-globin gene family. *Mol Biol Evol* **25**: 591–602.

Hoffmann FG, Storz JF (2007). The α^{D}-globin gene originated via duplication of an embryonic α-like globin gene in the ancestor of tetrapod vertebrates. *Mol Biol Evol* **24**: 1982–1990.

Hoffmann FG, Storz JF, Gorr TA, Opazo JC (2010). Lineage-specific patterns of functional diversification in the α- and β-globin gene families of tetrapod vertebrates. *Mol Biol Evol* **27**: 1126–1138.

Huelsenbeck JP, Ronquist F (2001). MRBAYES: Bayesian inference of phylogeny. *Bioinformatics* **17**: 754–755.

Hughes AL (1994). The evolution of functionally novel proteins after gene duplication. *Proc Biol Sci* **256**: 119–124.

Innan H, Kondrashov F (2010). The evolution of gene duplications: classifying and distinguishing between models. *Nat Rev Genet* **11**: 97–108.

Kaessmann H (2010). Origins, evolution, and phenotypic impact of new genes. *Genome Res* **20**: 1313–1326.

Kosakovsky Pond SL, Frost SDW (2005). Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics* **21**: 2531–2533.

Kosakovsky Pond SL, Frost SDW, Muse SV (2005). HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21**: 676–679.

Librado P, Rozas J (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**: 1451–1452.

Lynch M, Conery JS (2003). The evolutionary demography of duplicate genes. *J Struct Funct Genomics* **3**: 35–44.

Lynch M, Conery JS (2000). The evolutionary fate and consequences of duplicate genes. *Science* **290**: 1151–1155.

Maquat LE (2004). Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. *Nat Rev Mol Cell Biol* **5**: 89–99.

Maquat LE (2005). Nonsense-mediated mRNA decay in mammals. *J Cell Sci* **118**: 1773–1776.

Nielsen R, Yang Z (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**: 929–936.

Ohno S (1970). *Evolution by Gene Duplication*. Springer Verlag: New York.

Pegueroles C, Laurie S, Alba MM (2013). Accelerated evolution after gene duplication: a time-dependent process affecting just one copy. *Mol Biol Evol* **30**: 1830–1842.

Robin GC, Russell RJ, Cutler DJ, Oakeshott JG (2000). The evolution of an alpha-esterase pseudogene inactivated in the *Drosophila melanogaster* lineage. *Mol Biol Evol* **17**: 563–575.

Ronquist F, Huelsenbeck JP (2003). MRBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**: 1572–1574.

Sawyer SA (1989). Statistical tests for detecting gene conversion. *Mol Biol Evol* **6**: 526–538.

Sawyer SA (1999). GENECONV: A computer package for the statistical detection of gene conversion. Distributed by the author, Department of Mathematics, Washington University in St. Louis, available at http://www.math.wustl.edu/~sawyer.

Storz JF (2007). Hemoglobin function and physiological adaptation to hypoxia in high-altitude mammals. *J Mammal* **88**: 24–31.

Storz JF (2009). Genome evolution: gene duplication and the resolution of adaptive conflict. *Heredity* **102**: 99–100.

Storz JF, Baze M, Waite JL, Hoffmann FG, Opazo JC, Hayes JP (2007b). Complex signatures of selection and gene conversion in the duplicated globin genes of house mice. *Genetics* **177**: 481–500.

Storz JF, Hoffmann FG, Opazo JC, Moriyama H (2008). Adaptive functional divergence among triplicated α-globin genes in rodents. *Genetics* **178**: 1623–1638.

Storz JF, Opazo JC, Hoffmann FG (2011). Phylogenetic diversification of the globin gene superfamily in chordates. *IUBMB Life* **63**: 313–322.

Storz JF, Opazo JC, Hoffmann FG (2013). Gene duplication, genome duplication, and the functional diversification of vertebrate globins. *Mol Phylogenet Evol* **66**: 469–478.

Storz JF, Runck AM, Moriyama H, Weber RE, Fago A (2010). Genetic differences in hemoglobin function between highland and lowland deer mice. *J Exp Biol* **213**: 565–2574.

Storz JF, Sabatino SJ, Hoffmann FG, Gering EJ, Moriyama H, Ferrand N *et al.* (2007a). The molecular basis of high-altitude adaptation in deer mice. *PLoS Genet* **3**: 448–459.

Swofford DL (2003). *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)*. Version 4.0b10 Sinauer Associates: Sunderland, Massachusetts.

Tautz D, Domazet-Lošo T (2011). The evolutionary origin of orphan genes. *Nat Rev Genet* **12**: 692–702.

Tesakov AS, Lebedev VS, Bannikova AA, Abramson NI (2010). *Clethrionomys* Tilesius, 1850 is the valid generic name for red-backed voles and *Myodes* Pallas, 1811 is a junior synonym of *Lemmus* Link, 1795. *Russian J Theriol* **9**: 83–86.

Thein SL (2004). Genetic insights into the clinical diversity of beta thalassaemia. *Br J Haematol* **124**: 264–274.

Tufarelli C, Hardison R, Miller W, Hughes J, Clark K, Ventress N *et al.* (2004). Comparative analysis of the alpha-like globin clusters in mouse, rat, and human chromosomes indicates a mechanism underlying breaks in conserved synteny. *Genome Res* **14**: 623–630.

Wołk E (1983). Ontogenetic changes in the hemoglobin of the bank vole. *Acta Theriol* **28**: 387–396.
Yang Z (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13**: 555–556.
Yang Z (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol* **15**: 568–573.
Yang Z (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**: 1586–1591.
Yang Z, Wong WSW, Nielsen R (2005). Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol Biol Evol* **22**: 1107–1118.
Yingzhong Y, Yue C, Guoen J, Zhenzhong B, Lan M, Haixia Y *et al*. (2007). Molecular cloning and characterization of hemoglobin alpha and beta chains from plateau pika (*Ochotona curzoniae*) living at high altitude. *J Biochem Mol Biol* **40**: 426–431.
Zhang J (2003). Evolution by gene duplication: an update. *Trends Ecol Evol* **18**: 292–298.
Zhang J, Nielsen R, Yang Z (2005). Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol* **22**: 2472–2479.
Zwickl D (2006). *Genetic Algorithm Approaches for the Phylogenetic Analysis of Large Biological Sequence Datasets Under the Maximum Likelihood Criterion*. PhD Thesis. University of Texas at Austin, Austin, TX.

Supplementary Information accompanies this paper on Heredity website (http://www.nature.com/hdy)