Naked but Not *Hairless*: The Pitfalls of Analyses of Molecular Adaptation Based on Few Genome Sequence Comparisons

Frédéric Delsuc* and Marie-Ka Tilak

Institut des Sciences de l'Evolution, UMR5554, CNRS, IRD, Université de Montpellier, France

*Corresponding author: E-mail: frederic.delsuc@univ-montp2.fr.

Accepted: February 16, 2015

Data deposition: The newly obtained sequences have been deposited in the European Nucleotide Archive under accession numbers LN680723–LN680736.

Abstract

The naked mole-rat (*Heterocephalus glaber*) is the only rodent species that naturally lacks fur. Genome sequencing of this atypical rodent species recently shed light on a number of its morphological and physiological adaptations. More specifically, its hairless phenotype has been traced back to a single amino acid change (C397W) in the hair growth associated (HR) protein (or Hairless). By considering the available species diversity, we show that this specific position is in fact variable across mammals, including in the horse that was misleadingly reported to have the ancestral Cysteine. Moreover, by sequencing the corresponding *HR* exon in additional rodent species, we demonstrate that the C397W substitution is actually not a peculiarity of the naked mole-rat. Instead, this specific amino acid substitution is present in all hystricognath rodents investigated, which are all fully furred, including the naked mole-rat closest relative, the Damaraland mole-rat (*Fukomys damarensis*). Overall, we found no statistical correlation between amino acid changes at position 397 of the HR protein and reduced pilosity across the mammalian phylogeny. This demonstrates that this single amino acid change does not explain the naked mole-rat hairless phenotype. Our case study calls for caution before making strong claims regarding the molecular basis of phenotypic adaptation based on the screening of specific amino acid substitutions using only few model species in genome sequence comparisons. It also exposes the more general problem of the dilution of essential information in the supplementary material of genome papers thereby increasing the probability that misleading results will escape the scrutiny of editors, reviewers, and ultimately readers.

Key words: molecular evolution, adaptation, phylogenetics, mammals, rodents, *hairless*, naked mole-rat.

Introduction

Whole-genome sequencing offers unprecedented opportunity to reveal the genetic determinants of phenotypical characteristics of species. The availability of an increasing number of genome sequences allows conducting genome-wide scans of selection aimed at detecting the molecular footprints of adaptation (Heliconius Genome Consortium 2012; Qiu et al. 2012; Castoe et al. 2013; Schartl et al. 2013; Zhang et al. 2013; Fang, Nevo, et al. 2014; Fang, Seim, et al. 2014; Yim et al. 2014). In practice, this is often achieved by searching for lineage-specific amino acid changes that might cause functional changes in proteins encoded by candidate genes for a particular trait (e.g., Zhao et al. 2013). However, such a strategy can easily be misleading when only a limited number of sequences are used in the comparisons (Liu 2014). The publication of the naked mole-rat genome (Kim et al. 2011) provides a good illustration of both the utility and the potential drawbacks of this approach.

The rationale behind sequencing the genome of the naked mole-rat (*Heterocephalus glaber*) was the uniqueness of this eusocial rodent species with exceptional longevity, which serves as a model for aging, cancer, and pain resistance (Gorbunova et al. 2014; Keane et al. 2014). Mole-rats have actually evolved convergently in two distinct rodent clades with African mole-rats (Bathyergidae) belonging to Ctenohystrica and spalacids (Spalacidae) that are part of the mouse-related clade (Blanga-Kanfi et al. 2009). The genomes of the Damaraland mole-rat (*Fukomys damarensis*; Bathyergidae) and the blind mole-rat (*Nannospalax galili*; Spalacidae) have recently been sequenced to study

convergent genomic evolution in these ecologically and phenotypically similar species (Fang, Nevo, et al. 2014; Fang, Seim, et al. 2014). The naked mole-rat, which was long considered to be part of Bathyergidae, has recently been reclassified in its own family (Heterocephalidae) to better reflect its ancient divergence from other African mole-rats (more than 30 Ma) and its distinctive morphology (Patterson and Upham 2014).

The analysis of the naked mole-rat genome sequence has undoubtedly shed light on a number of genetic characteristics that might relate to its numerous peculiar physiological adaptations (Kim et al. 2011). Among the major insights was the finding that a single amino acid substitution was responsible for its unique hairless phenotype. Indeed, as its name implies, the naked mole-rat is the only wild rodent species that naturally lacks fur. The authors have linked this hairlessness to the presence of a single amino acid substitution in the nuclear receptor corepressor hair growth associated (HR) protein, also called Hairless. The *HR* gene is present as a single-copy ortholog in most currently available mammalian genomes with potential paralogs being only found in the gibbon genome (Ensembl Release 78). Kim et al. (2011) reported that the naked mole-rat HR sequence is substantially divergent from the other mammals investigated, and that it possesses a Tryptophan (Trp/W) at position 397 of the HR protein, whereas the other mammalian sequences compared exhibit a Cysteine (Cys/C). Because mutations at this particular codon of the *HR* gene are known to cause hair loss in mice, rat, and men (Panteleyev et al. 1998; Thompson 2009), they concluded that this specific amino acid substitution was likely responsible for the naked mole-rat hairless phenotype. The molecular evidence for such a conclusion takes the form of a multiple amino acid sequence alignment corresponding to exon 2 of the *HR* gene in rodents (exon 3 in human) presented as supplementary figure S24*C* of Kim et al. (2011). However, as we will show, these data are misleading in many aspects.

Results and Discussion

First, the authors chose to include only ten additional mammalian species in their HR comparative alignment even tough 32 orthologs of *HR* where available at the time in databases such as OrthoMaM v6 (Ranwez et al. 2007). Mining the different sequence and genome databases, we were able to retrieved 92 mammalian orthologous sequences for this exon (fig. 1). If the Cysteine observed at position 397 of the HR protein is likely to be ancestral in mammals because it is also found in monotremes and marsupials, this amino acid is in fact not conserved across placentals. Indeed, both the three camel species (genus *Camelus*) and the horse (*Equus caballus*) exhibit a Phenylalanine (Phe/F) at this position, resulting from two convergent nonsynonymous G->T substitutions at the second position of the corresponding codon. Contrary to what is shown in the alignment of supplementary figure S24*C* of Kim et al. (2011), the horse HR protein sequence they used (NCBI accession number: XP_001490941) actually has an F instead of a C at position 397 (supplementary fig. S1, Supplementary Material online). In fact, we verified that all available HR horse sequences have an F at this position (fig. 2) including an extinct 700,000 year-old individual (Orlando et al. 2013). The donkey (*Equus asinus*) also shows an F at this position whereas the rhino, which is the only other available perissodactyl representative, has the ancestral Cysteine. We thus come to the conclusion that this position has been mistakenly edited in the alignment of Kim et al. (2011).

Second, the guinea pig (*Cavia porcellus*), the closest relative of the naked mole-rat for which annotated genomic data were available at the time of the naked mole-rat genome publication, was not included in the HR protein sequence comparison reported by Kim et al. (2011). The absence of the guinea pig HR sequence in their alignment is surprising given that sequences of this species have been included in analyses of other 42,399 candidate genes in the same study (e.g., *TERF1* in their supplementary fig. S10). However, as in the naked mole-rat, the guinea pig HR sequence also has a W at position 397 (fig. 1). This suggests that this particular amino acid substitution is a shared characteristic of *Cavia* and *Heterocephalus* that both belong to Ctenohystrica, which includes Ctenodactylids (gundis) and Hystricognaths (e.g., *Cavia*) (Blanga-Kanfi et al. 2009).

By sequencing exons 1 and 2 of the *HR* gene (corresponding to exons 2 an 3 in human) in additional ctenohystrican representatives, we demonstrate that the presence of a W at position 397 is in fact a synapomorphy for hystricognaths (fig. 1). Indeed, the C397W substitution, which results from a nonsynonymous C->G nucleotide substitution at the third codon position (fig. 3), most likely occurred in the ancestral lineage of hystricognaths because the W is shared by all representative species investigated, whereas their closest gundi relative (*Ctenodactylus vali*) possesses the ancestral Cysteine. The closely related Damaraland mole-rat (*F. damarensis*; Bathyergidae), which is fully furred, also arbors the W found in the naked mole-rat. Conversely, all available nonhystricognath rodents, including the convergently evolved blind mole-rat (*N. galili*; Spalacidae), have the ancestral Cysteine. Because all these rodent species have fur, we conclude that this particular substitution in HR does not explain the hairless phenotype of the naked mole-rat. Among the 99 sequences investigated, the ancestral Cysteine residue is indeed fixed in the vast majority, including mammalian species with reduced pilosity such as aquatic cetaceans (Chen et al. 2013), the manatee (*Trichechus manatus*), and the walrus (*Odobenus rosmarus*), but also terrestrial species such as the nine-banded armadillo (*Dasypus novemcinctus*), the white rhino (*Ceratotherium simum*), and the African elephant (*Loxodonta africana*). Overall, we found no statistical correlation between amino acid changes at position 397 of HR and

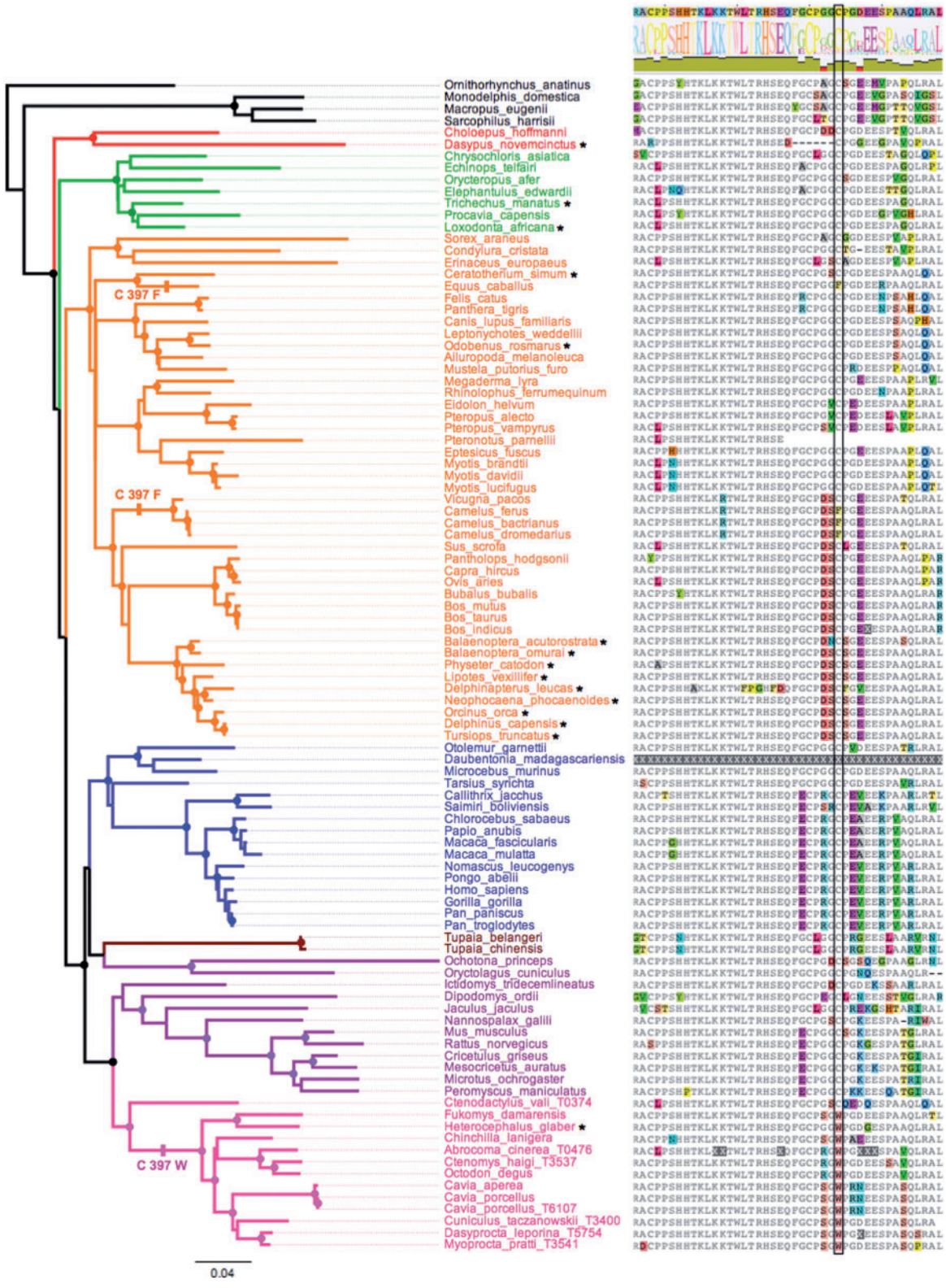

Fig. 1.—Maximum likelihood tree obtained from the concatenation of *HR* exons 1 and 2 for 99 mammals and partial amino acid alignment of exon 2 including position 397 (boxed). The three amino acid substitutions inferred at site 397 are mapped on the corresponding branches of the phylogenetic tree. Bullets indicate nodes with bootstrap percentage >90 and Bayesian posterior probability >0.95. Species are colored according to the placental clades to which they belong: Xenarthra (red), Afrotheria (green), Laurasiatheria (orange), Primates (blue), Scandentia (brown), Glires (purple), and Ctenohystrica (pink). Non-furred species are indicated by a star (*). Note that the ancestral therian and placental branches have been reduced by a factor ten to improve visibility of placental relationships. Sequence logos and the 50% majority-rule consensus are indicated as displayed in Geneious.

reduced pilosity across the mammalian phylogeny ($D = 0.0167$, $P = 0.68$; $M = 0.0028$, $P = 0.79$).

We also investigated whether the naked mole-rat *HR* gene sequence is really divergent from other mammals as suggested by Kim et al. (2011). Based on branch lengths of the maximum likelihood tree obtained from the concatenation of the two exons at the nucleotide level (fig. 1), the naked mole-rat does not appear to be especially fast evolving compare to other rodents. The *HR* gene evolves at sensibly the same pace in the two African mole-rats (*Heterocephalus* and *Fukomys*) and the convergently evolved spalacid (*Nannospalax*). The apparent sequence divergence of the naked mole-rat HR sequence observed by Kim et al. (2011) was thus likely an effect of their poor species sampling. Moreover, under the hypothesis that the *HR* gene is really involved in the naked mole-rat hairless phenotype, it should be possible to detect changes in selective constraints in this particular species. However, the *HR* gene was not identified as positively selected in recent genome-wide analyses of selection comparing *Heterocephalus* and *Fukomys* (Fang, Seim, et al. 2014).

Conclusions

This case study illustrates the pitfalls of the systematic quest for molecular adaptation by simply looking at point amino acid changes in candidate proteins coupled with the use of only few genome sequences. We suspect that numerous similar cases of false positive results will be revealed once a greater taxonomic diversity of genomes becomes available, as already shown by Liu (2014) for the *p53* gene. In mammals, for which a large number of genome sequences are already available, there is no reason not to use this phylogenetic diversity for empowering comparative evolutionary analyses aimed at detecting molecular adaptation. Our results also expose the more general problem of the dilution of essential information and figures in the ever-growing supplementary material of genome papers, which are often published in high profile journals where space is reduced. As shown here, such a practice significantly increases the probability that misleading results will escape the scrutiny of editors, reviewers, and ultimately readers.

Materials and Methods

Rodent tissue samples came from the Mammalian Tissue Collection of the Institut des Sciences de l'Evolution de Montpellier (Catzeflis 1991). We selected the following Ctenohystrica representatives: *Abrocoma cinerea* (T-0476), *C. porcellus* (T-6107), *C. vali* (T-0374), *Ctenomys haigi* (T-3537), *Cuniculus taczanowskii* (T-3400), *Dasyprocta leporina* (T-5754), and *Myoprocta pratti* (T-3541). DNA extractions were performed from tissues preserved in 95% ethanol using the DNeasy Blood & Tissue kit (QIAGEN) following manufacturer's instructions. We then polymerase chain reaction (PCR) targeted a 562 bp fragment of rodent HR exon 1 using the primer pair HR_ex2HysF (5′-GCCCAGCTTCCTGAAGGACAC-3′)/HR_ex2HysR (5′-CTTGCTGCCTAGGCYGAAGGC-3′) and a 584 bp fragment of exon 2 with primer pair HR_ex3HysF (5′-CTCAGGCTGGCAAAGGAGCC-3′)/HR_ex3HysR (5′-CTGCCTGC

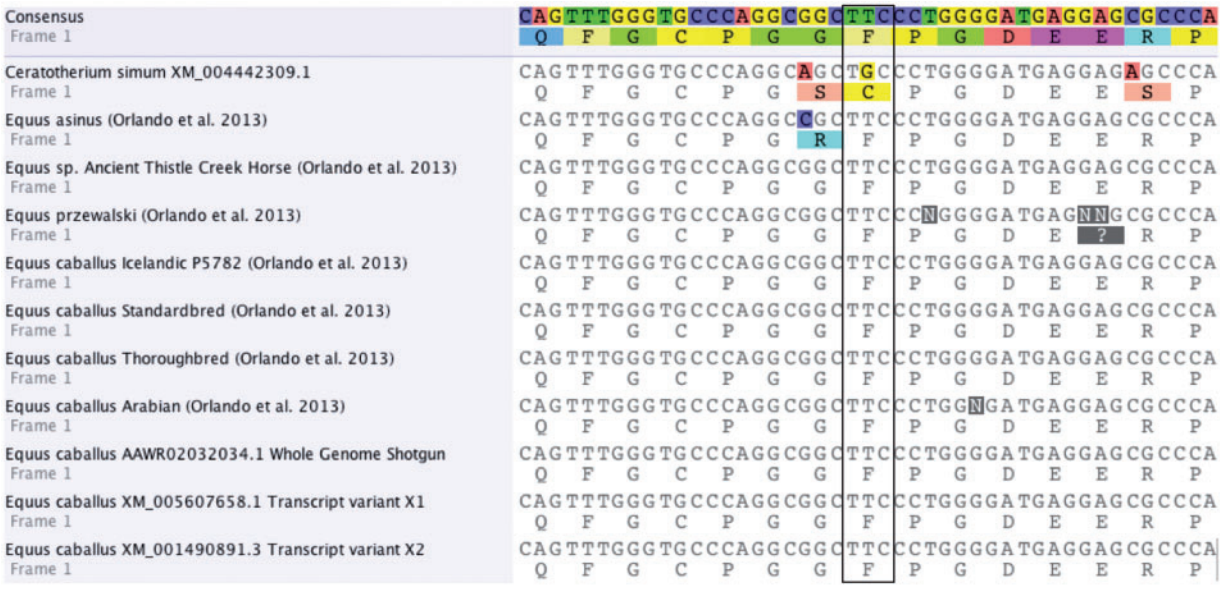

Fig. 2.—Partial multiple sequence alignment of *HR* exon 2 codons and corresponding amino acid translation for 11 perissodactyls including a diversity of extant and extinct equids. The codon corresponding to amino acid position 397 is boxed. The 50% majority-rule consensus is indicated as displayed in Geneious.

YCTCTTCAGGG-3′). The following PCR conditions were used for the two exons: 95 °C for 4 min (initial denaturation), followed by 30 cycles at 95 °C for 20 s (denaturation), 58 °C (exon 1) and 60 °C (exon 2) for 30 s (hybridization), 72 °C for 30 s (extension), and a final extension step at 72 °C for 10 min. PCR products were then purified with magnetic beads (Agencourt AMpure XP). Purified amplicons were Sanger sequenced on both strands using the PCR primers with the Big Dye Terminator V3.1 Kit (Applied Biosystem) on an Applied ABI Prism 3130XL automated sequencer. The newly obtained sequences have been deposited in the European Nucleotide Archive under accession numbers LN680723–LN680736.

Complete mammalian HR CDSs and HR exons 1 and 2 were extracted from the OrthoMaM v8 database (Douzery et al. 2014). These alignments were subsequently enriched by annotated sequences harvested from ongoing mammalian genome projects available as Whole Genome Shotgun assemblies in GenBank. A complete CDS data set was first assembled for 17 species of Glires. Newly obtained sequences for HR exons 1 and 2 were added to available orthologous sequences leading to a total of 99 sequences. Concatenations of HR exons 1 and 2 were built for 99 mammals and 25 species of Glires, respectively. All data sets have been aligned using the program MACSE (Ranwez et al. 2011) with default parameters, which allows conserving the coding frame. Ambiguously aligned codons were then excluded using the Gblocks server (Castresana 2000) with default relaxed parameters. All data sets are available upon request.

Maximum likelihood phylogenetic trees were inferred from nucleotide sequences for the exon data sets using the PhyML

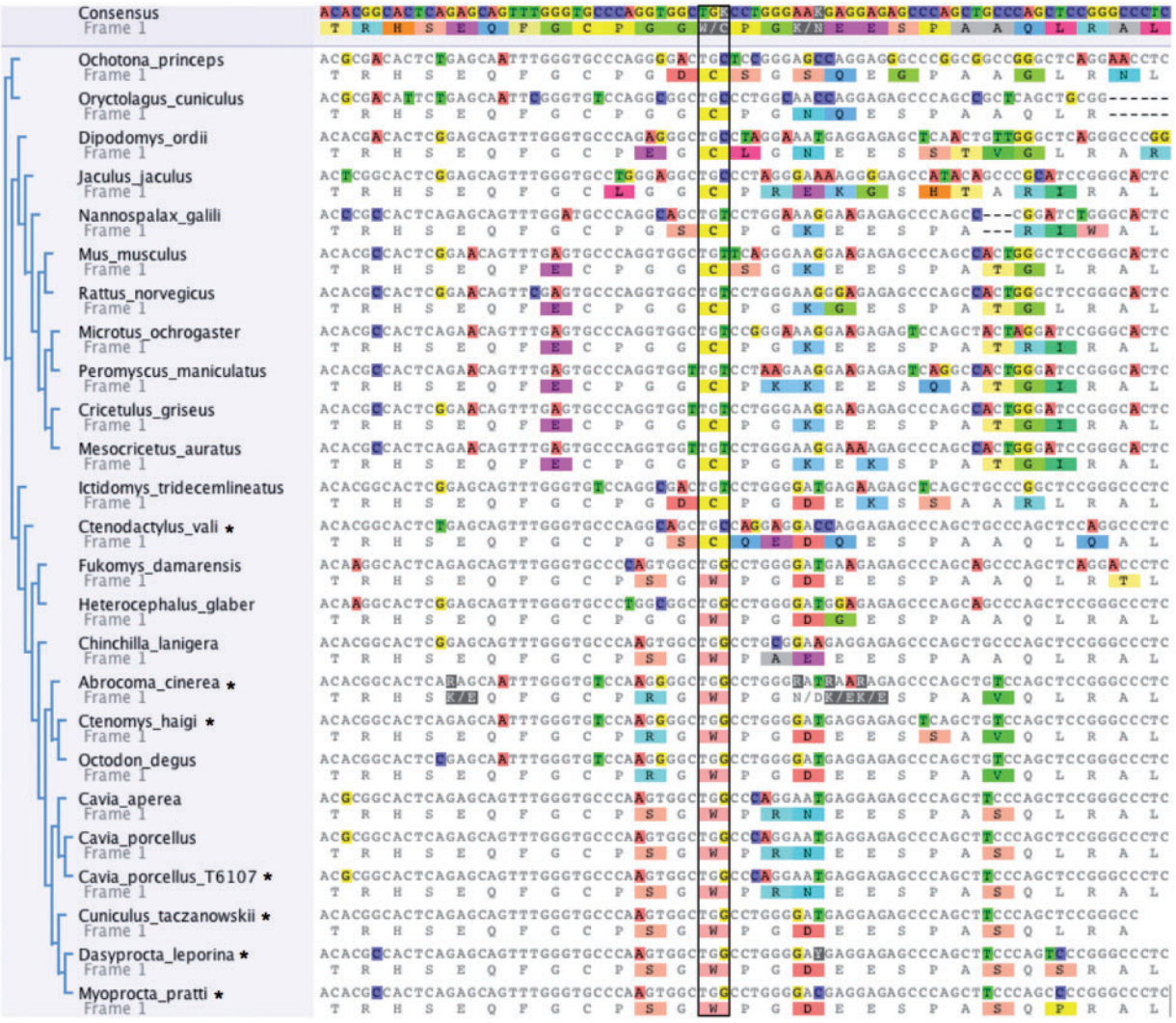

Fig. 3.—Partial multiple sequence alignment of *HR* exon 2 codon and corresponding amino acid translation for 25 Glires including seven new caviomorph sequences (*). The codon corresponding to amino acid position 397 is boxed. The 50% majority-rule consensus is indicated as displayed in Geneious.

3.0 (Guindon et al. 2009) plugin of Geneious R7 (Kearse et al. 2012) under the GTR+G8 model using SPR branch swapping on a BioNJ starting tree and 100 booststrap replications. Bayesian phylogenetic inference under a mixed model was conducted using the MPI version of MrBayes 3.2.3 (Ronquist et al. 2012). Separate GTR+G8 models were used for each codon position with parameters unlinked across partitions. Two independent runs of four incrementally heated MCMCMC starting from a random tree were performed. MCMCMC were run for 1,000,000 generations with trees and associated model parameters being sampled every 1,000 generations. The initial 250 trees in each run were discarded as burn-in samples after convergence checking. The 50% majority-rule Bayesian consensus tree and the associated posterior probabilities were computed from the 1,500 combined trees sampled in the two independent runs.

The correlation between amino acid changes at position 397 of HR and reduced pilosity was tested by calculating the *D* and *M* association statistics (Huelsenbeck et al. 2003) as implemented in SIMMAP 1.5 (Bollback 2006). To account for phylogenetic uncertainty, we used 15 trees subsampled from the previous posterior sample of 1,500 trees obtained with MrBayes. Amino acid changes were recoded as a multistate character ($C=0$; $W=1$; $F=2$) and pilosity was coded as a binary character (0 = fully furred; 1 = reduced pilosity; see fig. 1).

Supplementary Material

Supplementary figure S1 is available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors like to thank François Catzeflis, Patrick Gouat, Christopher J. Bonar, Gatine Yapok, James L. Patton, and Yuri Leite (Museum of Vertebrate Zoology, Berkeley, USA) for access to rodent tissue samples, Maeva Orliac for pointing out the HR gene, Ludovic Orlando and Mikkel Schubert for facilitating access to horse genomic data, Nicolas Galtier and Benoit Nabholz for fruitful discussions, and two anonymous referees for valuable comments. They also thank the sequencing centers that make their mammalian genome sequence data available. Sequence data were produced and analyzed through technical facilities of the Plateforme de Génomique Environnementale and the Montpellier Bioinformatics Biodiversity platform of the Labex CeMEB (Centre Méditerranéen de l'Environnement et de la Biodiversité). This work was supported by the Centre National de la Recherche Scientifique (CNRS). This is contribution ISEM 2015-014 of the Institut des Sciences de l'Evolution de Montpellier.

Literature Cited

Blanga-Kanfi S, et al. 2009. Rodent phylogeny revised: analysis of six nuclear genes from all major rodent clades. BMC Evol Biol. 9:71.

Bollback JP. 2006. SIMMAP: stochastic character mapping of discrete traits on phylogenies. BMC Bioinformatics 7:88.

Castoe TA, et al. 2013. The Burmese python genome reveals the molecular basis for extreme adaptation in snakes. Proc Natl Acad Sci U S A. 110: 20645–20650.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 17: 540–552.

Catzeflis FM. 1991. Animal tissue collections for molecular genetics and systematics. Trends Ecol Evol. 6:168.

Chen Z, Wang Z, Xu S, Zhou K, Yang G. 2013. Characterization of hairless (Hr) and FGF5 genes provides insights into the molecular basis of hair loss in cetaceans. BMC Evol Biol. 13:34.

Douzery EJP, et al. 2014. OrthoMaM v8: a database of orthologous exons and coding sequences for comparative genomics in mammals. Mol Biol Evol. 31:1923–1928.

Fang X, Nevo E, et al. 2014. Genome-wide adaptive complexes to underground stresses in blind mole-rats *Spalax*. Nat Commun. 5: 3966.

Fang X, Seim I, et al. 2014. Adaptations to a subterranean environment and longevity revealed by the analysis of mole-rat genomes. Cell Rep. 8:1354–1364.

Gorbunova V, et al. 2014. Comparative genetics of longevity and cancer: insights from long-lived rodents. Nat Rev Genet. 15:531–540.

Guindon S, Delsuc F, Dufayard JF, Gascuel O. 2009. Estimating maximum likelihood phylogenies with PhyML. Methods Mol Biol. 537: 113–137.

Heliconius Genome Consortium. 2012. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature 487: 94–98.

Huelsenbeck JP, Nielsen R, Bollback JP. 2003. Stochastic mapping of morphological characters. Syst Biol. 52:131–158.

Keane M, et al. 2014. The Naked Mole-rat Genome Resource: facilitating analyses of cancer and longevity-related adaptations. Bioinformatics 30:3558–3560.

Kearse M, et al. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28:1647–1649.

Kim EB, et al. 2011. Genome sequencing reveals insights into physiology and longevity of the naked mole-rat. Nature 479:223–227.

Liu Z. 2014. Codon 104 of p53 is not an adaptively selected site for extreme environments in mammals of the Tibet plateau. Proc Natl Acad Sci U S A. 111:E2357.

Orlando L, et al. 2013. Recalibrating *Equus* evolution using the genome sequence of an early Middle Pleistocene horse. Nature 499:74–78.

Panteleyev AA, Paus R, Ahmad W, Sundberg JP, Christiano AM. 1998. Molecular and functional aspects of the hairless (HR) gene in laboratory rodents and humans. Exp Dermatol. 7:249–267.

Patterson BD, Upham NS. 2014. A newly recognized family from the Horn of Africa, the Heterocephalidae (Rodentia: Ctenohystrica). Zool J Linn Soc. 172:942–963.

Qiu Q, et al. 2012. The yak genome and adaptation to life at high altitude. Nat Genet. 44:946–949.

Ranwez V, et al. 2007. OrthoMaM: a database of orthologous genomic markers for placental mammal phylogenetics. BMC Evol Biol. 7:241.

Ranwez V, Harispe S, Delsuc F, Douzery EJP. 2011. MACSE: multiple alignment of coding sequences accounting for frameshifts and stop codons. PLoS One 6:e22594.

Ronquist F, et al. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol. 61: 539–542.

Schartl M, et al. 2013. The genome of the platyfish, *Xiphophorus maculatus*, provides insights into evolutionary adaptation and several complex traits. Nat Genet. 45:567–572.

Thompson CC. 2009. Hairless is a nuclear receptor corepressor essential for skin function. Nucl Recept Sign. 7:e010.
Yim HS, et al. 2014. Minke whale genome and aquatic adaptation in cetaceans. Nat Genet. 46:88–92.
Zhang G, et al. 2013. Comparative analysis of bat genomes provides insight into the evolution of flight and immunity. Science 339:456–460.
Zhao Y, et al. 2013. Codon 104 variation of p53 gene provides adaptive apoptotic responses to extreme environments in mammals of the Tibet plateau. Proc Natl Acad Sci U S A. 110: 20639–20644.

Associate editor: Bill Martin