

Molecular evolution of the hyaluronan synthase 2 (*HAS2*) gene in mammals: implications for adaptations to the subterranean niche and cancer resistance.

S1. SUPPLEMENTARY METHODS

(a) Sampling, PCR and sequencing

Samples newly sequenced for this study and their geographic origins are outlined in Table S1. Genomic DNA was extracted from muscle using a standard protocol [1], and PCR amplification of exons 2 and 4 of the *HAS2* gene carried out using primers designed using Primer3 [2, 3] from conserved regions in multiple species alignments of published *HAS2* sequences. Exon 2 (627 bp; 209 amino acids): HAS2E1F (5'-TGC ATT GTG AGA GGT TTM TAT G-3') and HAS2E1R (5'- CTG WAC ATA RTC CAC RCT BCG-3'); exon 4 (930bp; 310 amino acids): HAS2E3F (5'-CAA RTA YGA YTC STG GAT YTC YTT-3') and HAS2E3R (5'- CAT ACR TCM AGC ACC ATG TC-3'). PCR reactions followed a standard protocol using 10 ng of DNA in a 50 μ l reaction containing Taq buffer, 1 Unit of Taq polymerase, 0.4 μ M of each primer and 200 μ M of dNTPs. After initially denaturing at 94°C for 2 min, 35 cycles of 94°C (30 s), 50°C (30 s), and 72°C (30 s) were carried out, followed by a final 72°C for 10 min. Negative controls were also included in each set of reactions. PCR products were purified using QIAquick PCR purification kits® (Qiagen, UK). Exons 1 and 3 were not sequenced as the former is non-transcribed, and the latter short (102 bp; 34 amino acids) and invariant in amino acid sequence across mammals. Sequencing was carried out in both directions using the PCR primers to obtain complementary partially or fully overlapping strands, using the Eurofins Genomics Value Read automated sequencing service (Eurofins Genomics, Ebersberg, Germany). In species where sequencing failed using the above primers, PCR products were cloned using a pGEM®-T Easy Vector System (Promega, UK), and positive clones sequenced using standard plasmid primers T7 and SP6.

(b) Phylogenetic analysis

In addition to *de novo* sequencing of DNA from the 13 species in Table S1, a further 57 representative mammalian *HAS2* sequences from all species available at the time of the study were retrieved from GenBank (Table 1). Sequences were aligned manually for analysis using Mesquite [4], and the species tree topology (e.g. Figures S1-3) used for selection analyses based on published studies [5, 6, 7]. Genetic distances were calculated and phylogenetic trees based on both nucleotide and amino acid sequences constructed using MEGA 6 [8].

(c) Testing for selection

Site models

We characterised the signature of selection acting along *HAS2* across all 70 mammal species included in our study. Site-wise ω values were estimated across all branches in the phylogeny for the alignment consisting of 519 codons run with the codeml package in PAMLv4.4 [9]. These ω values were assigned to predefined site classes according to each site model (e.g. M1a had two and M2a had three) [10, 11]. We estimated the mean ω of each site class, and the proportions of sites falling into each class. To test where and how ω varied among sites; three model comparisons were carried out. Firstly, we assessed whether ω varied among sites by comparing a model with a single free ω (M0) to one in which ω fell into three discrete classes (M3). Secondly, to test for positive selection, we compared model M1a (Nearly Neutral) in which site classes were neutral ($\omega=1$) and purifying ($0<\omega<1$) to model M2a (Positive Selection) that had a third site class corresponding to positive selection ($\omega>1$). For a second test of adaptation, we compared model M7 (beta) to M8 (beta & ω), in which the

latter had an additional site class in which ω could exceed one. Likelihood ratio tests (LRT) were used for all model comparisons.

Branch-site models

In order to determine if any particular sites along *HAS2* were under positive selection in two particular branches of interest we implemented branch-site models [12]. In our first test we set the naked mole-rat branch as the foreground branch of interest; in the second test the ancestral mole-rat + cane rat branch was set as the foreground branch of interest. In each branch-site test for positive selection the site-wise estimates of ω for the branch that has been designated as the foreground branch of interest are compared with estimates across the remaining background branches in the species phylogeny under model A. The model parameters are then compared with those of the null model A with a likelihood ratio test (LRT), and the significance of the model fit assessed by ChiSq statistic with one degree of freedom, with *P*-values <0.05 indicating the alternative model has a significantly better fit compared to the null.

Clade models

We also tested for evidence of divergent selection pressures acting on the focal clade of interest, corresponding to the Bathyergidae + cane rat, compared to outgroup taxa. In this test all branches of the Bathyergidae + cane rat clade, including the ancestral branch is set as the foreground clade and the remaining branches are the background clade (see Figure S1 for tree topology). We implemented the clade model C [13], in which the estimated averaged ω of the foreground clade is compared to that of the averaged ω of the background clade. Values estimated by each clade model were then compared with model M1a (nearly neutral) via a LRT with three (DF), again with *P*-values <0.05 indicating the alternative model has a significantly better fit compared to the null.

(d) Multivariate Analysis of Protein Polymorphism

Multivariate Analysis of Protein Polymorphism (MAPP) uses an amino acid sequence alignment to calculate the predicted impact of each potential substitution at each position. These predictions are based on a set of scales of physicochemical properties (hydropathy, polarity, charge, volume and free energy in alpha-helix and beta-strand conformation), for which each amino acid has a numeric value. *P*-value interpretations of the MAPP scores are then calculated, predicting the impact of each amino acid variant [14; http://mendel.stanford.edu/supplementarydata/stone_MAPP_2005]

S2. SUPPLEMENTARY RESULTS

(a) Phylogenetic analysis

A summary of pairwise genetic distances for the complete mammalian dataset of 70 species are presented in Table S1. Nucleotide uncorrected *p*-distances ranged from 0.001 (humans versus chimps) to 0.195 (Philippine tarsier versus opossum), reflecting phylogenetic proximity on the species tree (Figures S1-3). Amino acid differences ranged from zero among all anthropoid primates except the bonobo, where there is a unique phenylalanine (F) to leucine (L) substitution at site 398 (an extracellular region), to 0.153 (Cape elephant shrew versus opossum). Interestingly, the opossum consistently across all sites shows greater numbers of amino acid differences with eutherian mammals than the more divergent platypus, despite greater nucleotide distances in the latter (Table S1). Maximum likelihood trees were inferred by using the K2+G+I model (for nucleotide sequences) and the JTT+G model (for amino acid sequences). The nucleotide based tree recovered many of the main clades of the species tree (as in Figures S1-3) with bootstrap support of >90%, with anomalous groupings occurring as polytomies or clades with very low bootstrap values. The

amino acid based tree was not informative as only three clades were recovered with bootstrap support of >90%, one with the African mole-rats *excluding* the naked mole-rat (93%), another with four Afrotherians (Cape elephant shrew, aardvark, tenrec and golden mole; 98%) and the third with the bats from the Vespertilionidae family (93%).

(b) Selection analysis

Site models

Our site-wise analysis of *HAS2* sequences across 70 mammal species found no evidence of positive selection as both pair-wise tests for positive selection (M1a vs. M2a and M7 vs. M8) were not significant (LRT: $P=1.000$ and $P=0.427$). Instead, sites along *HAS2* were found to be under purifying selection with all three estimated ω categories of Model 3 (discrete) being <1.0 (LRT: M0 vs. M3 $P<0.0001$) (Table 1).

Branch-site models

There was no evidence of sites under positive selection along the naked mole-rat branch (LRT: $P=1.000$) or along the ancestral branch of the Bathyergidae + cane rat (LRT: $P=1.000$) in either of our branch-site tests.

Clade models

Clade models performed to test for divergent selection between the focal clade of Bathyergidae + cane rat compared to the remaining mammal species detected significant divergent selection (LRT: $P<0.001$). However, the estimated omega of both the foreground and background clades was <1.00 and, therefore, both clades were found to be under purifying selection ($FG\omega = 0.156$; $BG\omega = 0.126$).

Clade models performed to test for divergent selection between the focal clade of Bathyergidae + cane rat compared to the remaining Euarchontoglires species (Rodentia, Lagomorpha, Primates and Scandentia) detected significant divergent selection (LRT: $P<0.001$). However, the estimated omega of both the foreground and background clades was <1.00 and, therefore, both clades were found to be under purifying selection ($FG\omega = 0.215$; $BG\omega = 0.112$).

Clade models performed to test for divergent selection between the focal clade of Bathyergidae + cane rat compared to a background clade of four outgroup Hystricomorph species (porcupine, tuco-tuco, chinchilla and guinea pig) did not detect significant divergent selection (LRT: $P=1.000$). Therefore, the null model (M1a nearly neutral) was found to fit the data better, with the majority of sites (~ 0.99) found to be under purifying selection ($\omega = 0.020$).

(c) Multivariate Analysis of Protein Polymorphism (MAPP) analysis

A full summary of the numbers of substitutions per taxon that are predicted by MAPP analysis to have a significant impact on protein function is presented in Table S3. These values ranged from zero to a maximum of 21 significant substitutions in the Cape elephant shrew, closely followed by the opossum with 20 substitutions. Third ranking was a much lower value of 11, observed in the aardvark. In this context, the two significant substitutions of the naked mole-rat fall at the bottom end of the distribution. Amino acid residues that are crucial for *N*-acetylglucosaminyltransferase activity at sites 212 (aspartic acid), 314 (aspartic acid), 350 (glutamine), 353 (arginine) and 354 (tryptophan) [15] are fully conserved across all mammals in our dataset (See Figure S4).

S3. SUPPLEMENTARY TABLES

Table S1: Geographic origins for the samples and species sequenced for this study, together with maximum lifespan data (years) where known [16, 17]. See [18] for current nomenclature for *Heliophobius*.

Species	Common name	Order/Family	Location	Longevity
<i>Amblysomus hottentotus</i>	golden mole	Afrosoricida; Chrysochloridae	Glengarry, South Africa	?
<i>Thryonomys swinderianus</i>	cane rat	Rodentia; Thryonomyidae	Natal, South Africa	5.4
<i>Hystrix africaeaustralis</i>	Cape porcupine	Rodentia; Hystricidae	Captive, London Zoo	23
<i>Ctenomys perrensi</i>	tucu tucu	Rodentia; Ctenomyidae	Iberá, Argentina	?
<i>Talpa europaea</i>	European mole	Eulipotyphla; Talpidae	Norfolk, UK	?
<i>Tachyoryctes splendens</i>	East African root rat	Rodentia; Spalacidae	Kilimanjaro, Tanzania	?
<i>Fukomys zechi</i>	Ghana mole-rat	Rodentia; Bathyergidae	Atebubu, Ghana	?
<i>Fukomys damarensis</i>	Damaraland mole-rat	Rodentia; Bathyergidae	Hotazel, South Africa	15.5
<i>Georychus capensis</i>	Cape dune mole-rat	Rodentia; Bathyergidae	Darling, South Africa	<11
<i>Cryptomys hottentotus</i>	common mole-rat	Rodentia; Bathyergidae	Frankenwald, South Africa	<11
<i>Bathyergus janetta</i>	Namaqua dune mole-rat	Rodentia; Bathyergidae	Rondawel, South Africa	4-6?
<i>Bathyergus suillus</i>	dune mole-rat	Rodentia; Bathyergidae	Stilbaai, South Africa	4-6?
<i>Heliophobius kapiti</i>	silvery mole-rat	Rodentia; Bathyergidae	Athi Plains, Kenya	7

S4. SUPPLEMENTARY FIGURES

Figure S1: Species tree illustrating the character evolution at Site 177 of exon 2 of the *HAS2* gene, indicating gains and losses of alanine.

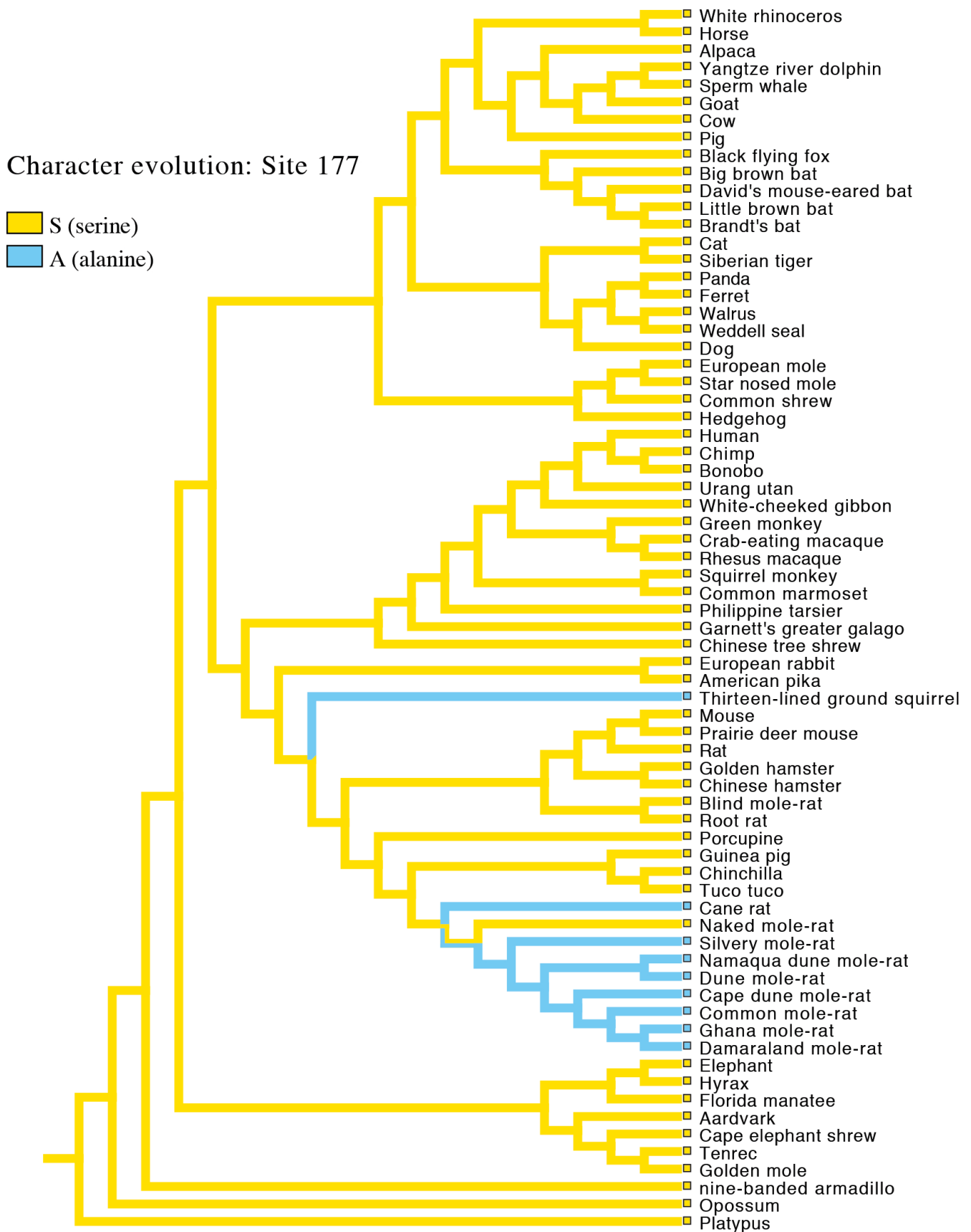


Figure S2: Species tree illustrating the character evolution at Site 178 of exon 2 of the *HAS2* gene, indicating gains and losses of serine.

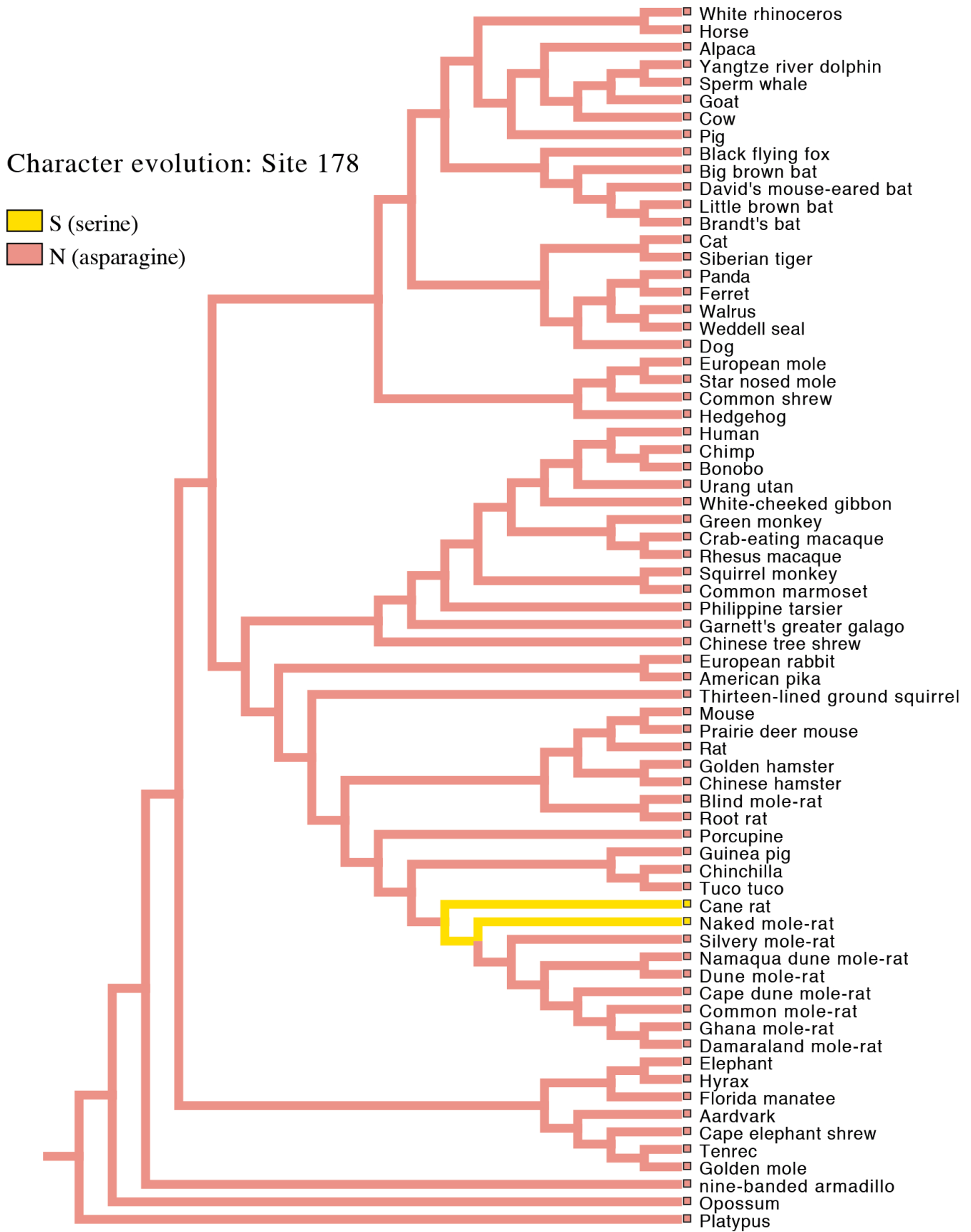


Figure S3: Species tree illustrating the character evolution at Site 301 of exon 4 of the *HAS2* gene, indicating gain of a serine substitution in the bathyergid clade.

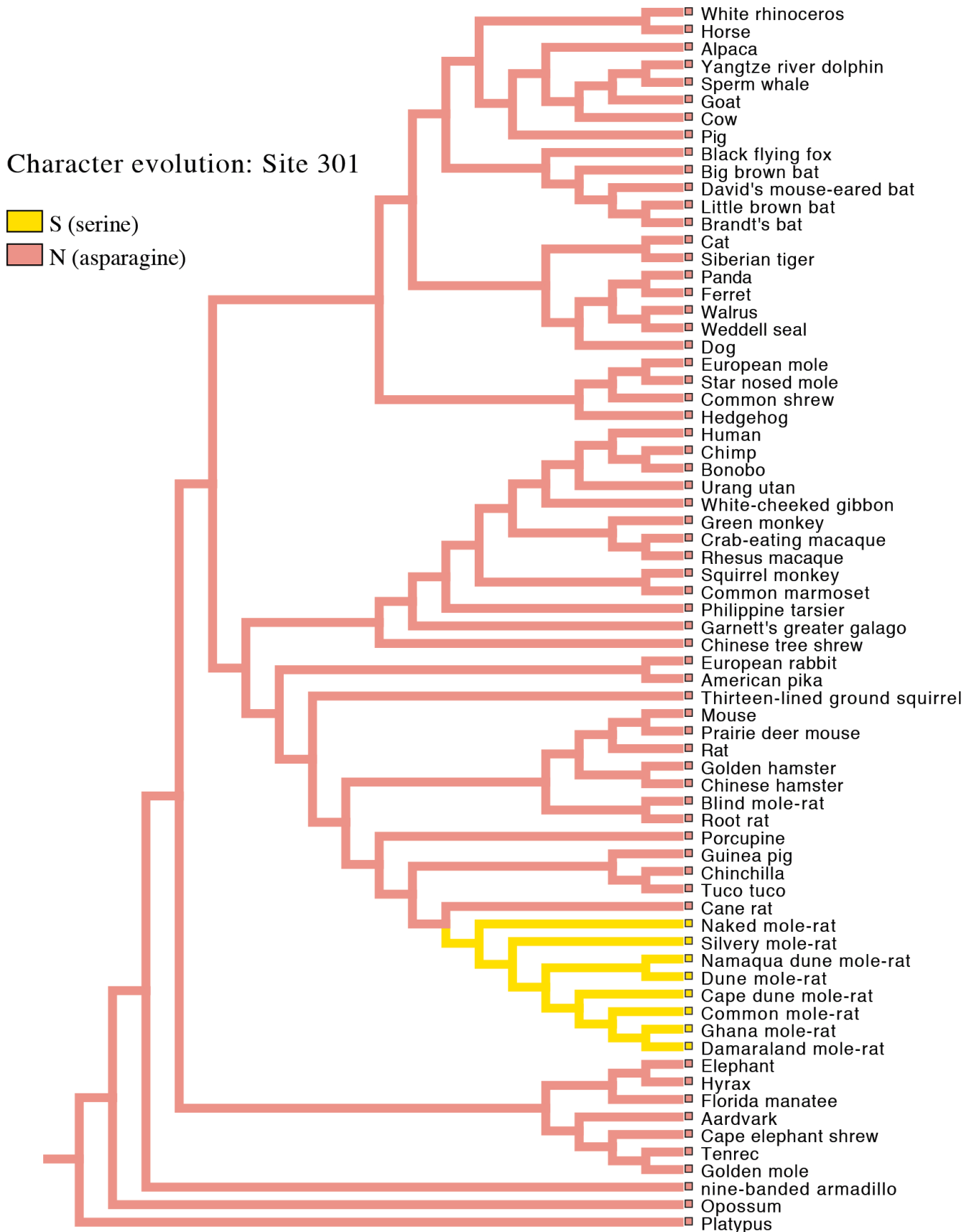
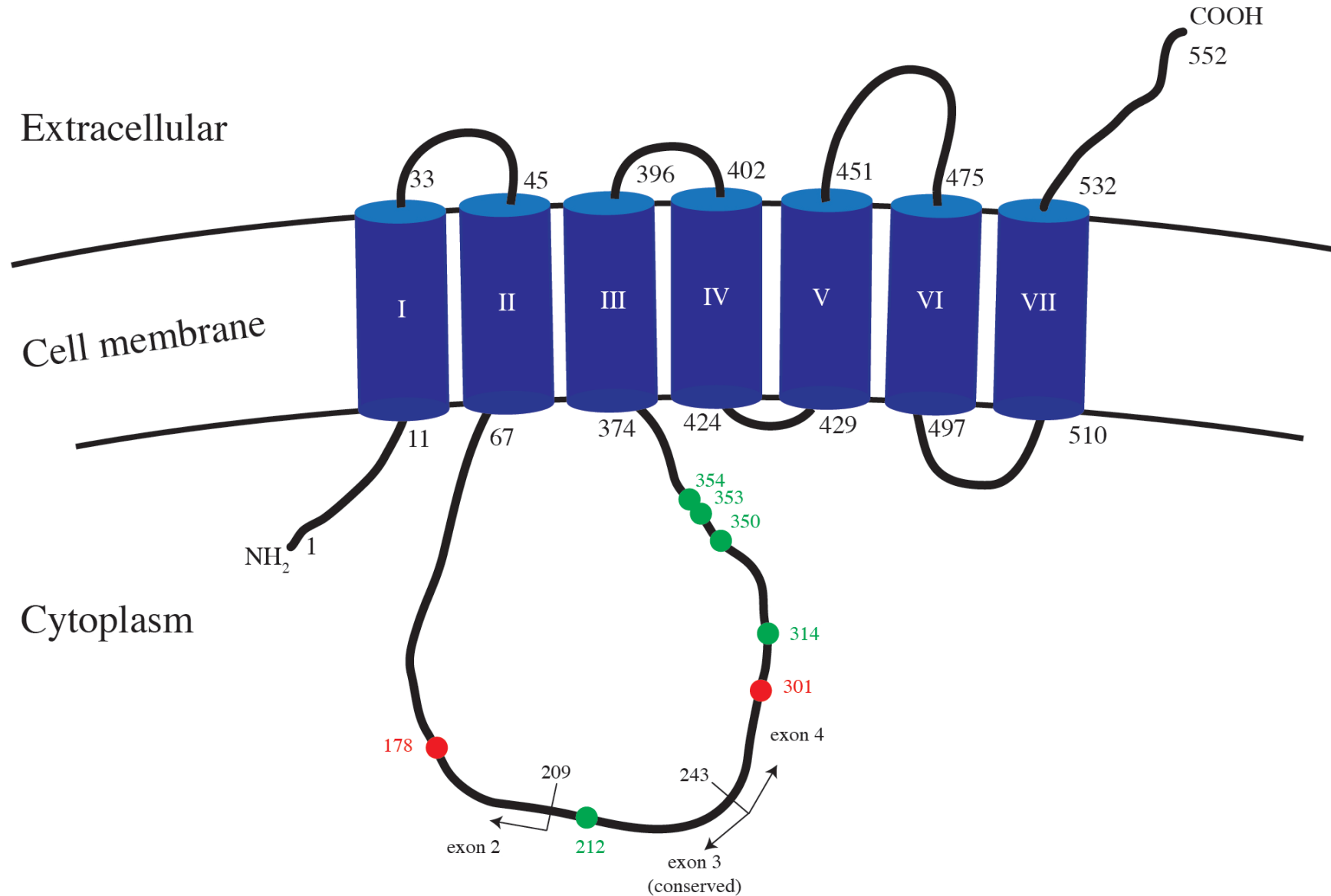


Figure S4: Schematic representation of the HAS2 molecule (predicted) with key sites (numbered), exons and domains marked. Coloured cylinders I to VII represent transmembrane helical domains; extracellular and cytoplasmic domains, with start and finish sites as indicated by black lines. Green dots and numbers correspond to sites thought to be crucial for N-acetylglucosaminyltransferase activity (Watanabe & Yamaguchi, 1996), red dots and numbers correspond to the two sites in the naked mole-rat implicated in the production of HMM-HA.



Supplementary references

1. Faulkes CG, Abbott DH, O'Brien HP, Lau L, Roy M, Wayne RK, Bruford MW. 1997 Micro- and macro-geographic genetic structure of colonies of naked mole-rats, *Heterocephalus glaber*. *Mol. Ecol.* **6**, 615–628.
2. Koressaar T, Remm M. 2007 Enhancements and modifications of primer design program Primer3. *Bioinformatics* **23**, 1289–91.
3. Untergrasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. 2012 Primer3 - new capabilities and interfaces. *Nucleic Acids Res.* **40**, e115.
4. Maddison WP, Maddison DR. 2014 Mesquite: a modular system for evolutionary analysis. Version 3.01 <http://mesquiteproject.org>
5. Faulkes CG, Verheyen E, Verheyen W, Jarvis JUM, Bennett NC. 2004 Phylogeographic patterns of speciation and genetic divergence in African mole-rats (Family Bathyergidae). *Mol. Ecol.* **13**, 613–629.
6. Blanga-Kanfi S, Miranda H, Penn O, Pupko T, DeBry R, Huchon D. 2009 Rodent phylogeny revised: analysis of six nuclear genes from all major rodent clades. *BMC Evol. Biol.* **9**, 71. (doi:10.1186/1471-2148-9-71)
7. Meredith RW, Janecka JE, Gatesy J, Ryder OA, Fisher CA, et al. 2011 Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification. *Science* **334**, 521–524. (doi: 10.1126/science.1211028)
8. Tamura K, Stecher G, Peterson D, Filipowski A, Kumar S. 2013 MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol.* **30**, 2725–2729.
9. Yang ZH. 2007 PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591.
10. Nielsen R, Yang ZH. 1998 Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**, 929–936.
11. Wong WSW, Yang Z, Goldman N, Nielsen R. 2004 Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics* **168**, 1041–1051.
12. Zhang JZ, Nielsen R, Yang ZH. 2005 Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.* **22**, 2472–2479.

13. Bielawski JP, ZH Yang. 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.
14. Stone EA, Siddow A. 2005 Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Res.* **15**, 978–986.
15. Watanabe K, Yamaguchi Y. 1996 Molecular identification of a putative human hyaluronan synthase. *J. Biol. Chem.* **271**, 22945–22948.
16. Weigl R. 2005 Longevity of Mammals in Captivity; from the Living Collections of the World. Kleine Senckenberg-Reihe 48: Stuttgart.
17. Tacutu R, Craig T, Budovsky T, Wuttke D, Lehmann G, Taranukha D, Costa J, Fraifeld VE, de Magalhães JP. 2013 Human Ageing Genomic Resources: Integrated databases and tools for the biology and genetics of ageing. *Nucleic Acids Res.* **41**, D102–D1033. (doi: 10.1093/nar/gks1155)
18. Faulkes CG, Bennett NC, Cotterill FPD, Stanley W, Mgone GF, Verheyen E. 2011 Phylogeography and cryptic diversity of the solitary-dwelling silvery mole-rat, genus *Heliophobius* (Family: Bathyergidae). *J. Zool. (Lond.)* **285**, 324–338. (doi:10.1111/j.1469-7998.2011.00863.x)