Adaptations to a subterranean environment and longevity revealed by the analysis of mole rat genomes

Xiaodong Fang[1,2,10], Inge Seim[3,4,10], Zhiyong Huang[1], Maxim V. Gerashchenko[3], Zhiqiang Xiong[1], Anton A. Turanov[3], Yabing Zhu[1], Alexei V. Lobanov[3], Dingding Fan[1], Sun Hee Yim[3], Xiaoming Yao[1], Siming Ma[3], Lan Yang[1], Sang-Goo Lee[4], Eun Bae Kim[4], Roderick T. Bronson[5], Radim Šumbera[6], Rochelle Buffenstein[7], Xin Zhou[1], Anders Krogh[2], Thomas J. Park[8], Guojie Zhang[1,2], Jun Wang[1,2,9,*], Vadim N. Gladyshev[3,4,*]

[1]BGI-Shenzhen, Shenzhen, 518083, China

[2]Department of Biology, University of Copenhagen, Copenhagen, DK-2200 Copenhagen N, Denmark

[3]Division of Genetics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, 02115, USA

[4]Department of Bioinspired Science, Ewha Womans University, Seoul, 120-750, South Korea

[5]Rodent Histopathology Laboratory, Harvard Medical School, Boston, MA 02115, USA

[6]University of South Bohemia, Faculty of Science, Ceske Budejovice, 37005, Czech Republic

[7]Department of Physiology and The Sam and Ann Barshop Institute for Longevity and Aging Studies, University of Texas Health Science Center, San Antonio, TX 78245, USA

[8]Department of Biological Sciences, University of Illinois at Chicago, Chicago, IL 60607, USA

[9]King Abdulaziz University, Jeddah, 21441, Saudi Arabia

[10]Co-first authors

[*]Correspondence: wangj@genomics.org.cn (J.W.), vgladyshev@rics.bwh.harvard.edu (V.N.G)

Supplemental Information

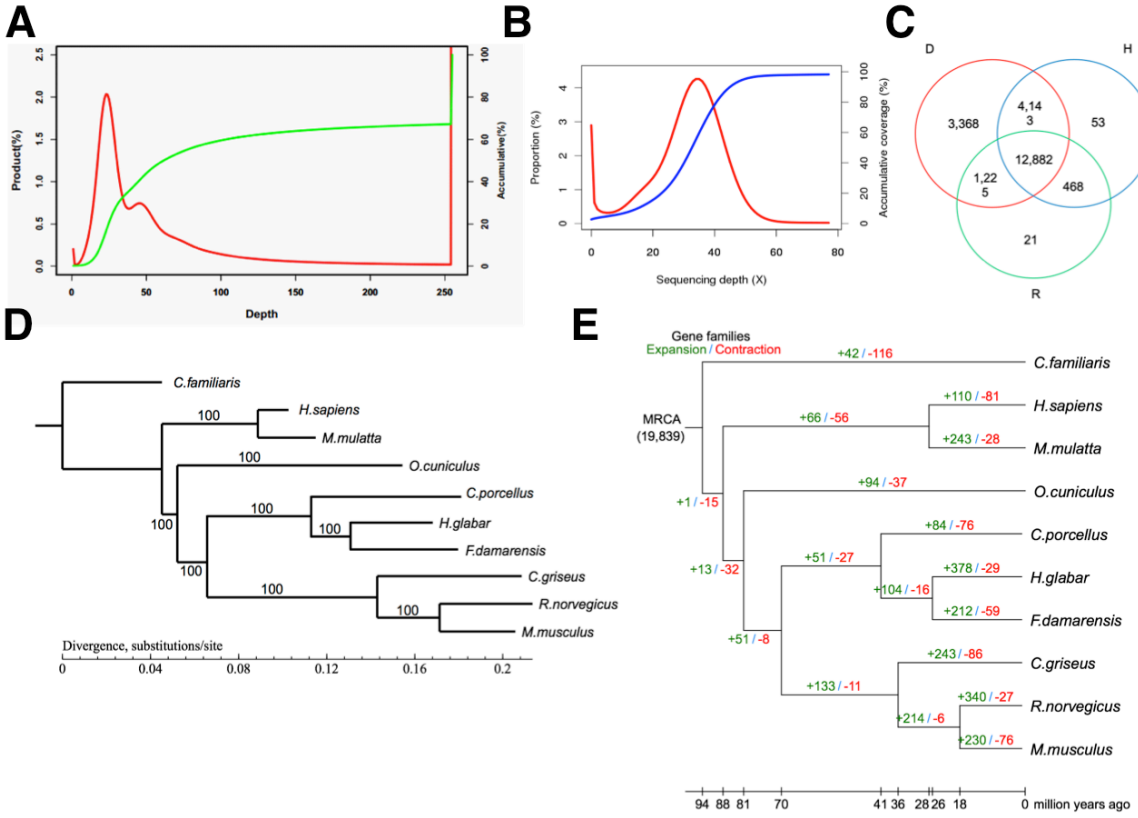


Figure S1, Related to Figure 1. Features of the DMR genome

(A) DMR genome size estimation using k-mer analysis. The red curve refers to the k-mer frequency distribution, and the green curve refers to the cumulative distribution of k-mer frequency. More than 25% high depth kmer(>255) may due to repeats in the genome. An unexpected peak at depth of 50 may result from segmental duplication or other repetitive elements. **(B)** Sequencing depth distribution of the DMR genome. High quality short insert size reads were mapped to the associated genome with an average depth of 34, and approximately 91.64% of the genome was covered by more than 10 reads. The red curve denotes the proportion of the genome in a given sequencing depth, and the blue curve shows the cumulative coverage of the genome. **(C)** Summary of evidence for DMR gene models from three types of gene source (D: *de novo* prediction; H: homology-based; R: RNA-seq). **(D)** Phylogenetic tree of 10 mammalian species. The tree was constructed with PhyML under an ml+JTT+gamma model. The bootstrap values were calculated based on 1,000 replicates. **(E)** Distribution of lost and newly evolved gene families in mammalian lineages. Numbers designate the number of gene families that have expanded (green) and contracted (red) since the split from the common ancestor. The most recent common ancestor (MRCA) has 19,839 gene families.

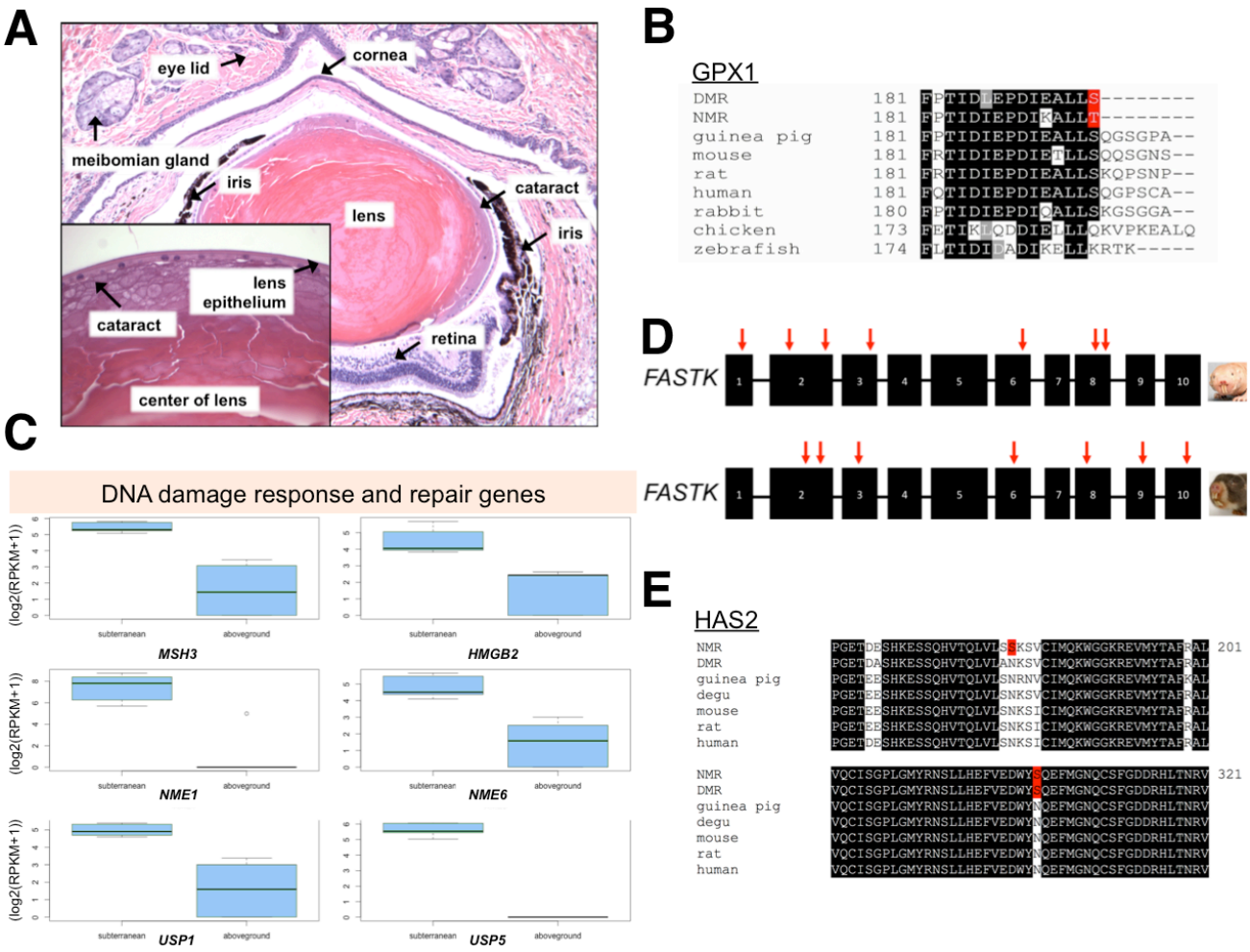


Figure S2, Related to Figures 2 and 3. Morphological and molecular changes of subterranean rodents

(A) NMR eye (x10, insert at x60). NMRs have a normal eye morphology with the exception of cataracts, which were identified in every specimen examined (n=4). **(B)** The gene encoding glutathione peroxidase 1 (*GPX1*) of NMR and DMR harbors a premature stop codon and results in a truncated protein. Sequence alignment of vertebrate GPx1 proteins (C-terminal regions) is shown, and the C-terminal amino acid residue of NMR and DMR is highlighted in red. **(C)** Expression pattern of DNA damage response and repair genes in subterranean and 'aboveground' (surface-dwelling) rodents. The selected genes include mutS homolog 3 (*MSH3*), high mobility group box 2 (*HMGB2*), NME/NM23 nucleoside diphosphate kinase 1 and 6 (*NME1* and *NME6*), and ubiquitin specific peptidase 1 and 5 (*USP1* and *USP5*). **(D)** Overview of inactivation events in *FASTK* of African mole rats. The exon structure of the human gene, which is identical to the mouse and rat ortholog, is shown. Boxes indicate exons and lines introns, while red arrows indicate inactivation events (insertions or deletions that change the frame, or stop point mutations resulting in premature termination of translation). **(E)** Sequence alignment of mammalian hyaluronan synthase 2 (HAS2). Amino acids 178 and 301 are highlighted in red. Identical residues are shaded in black.

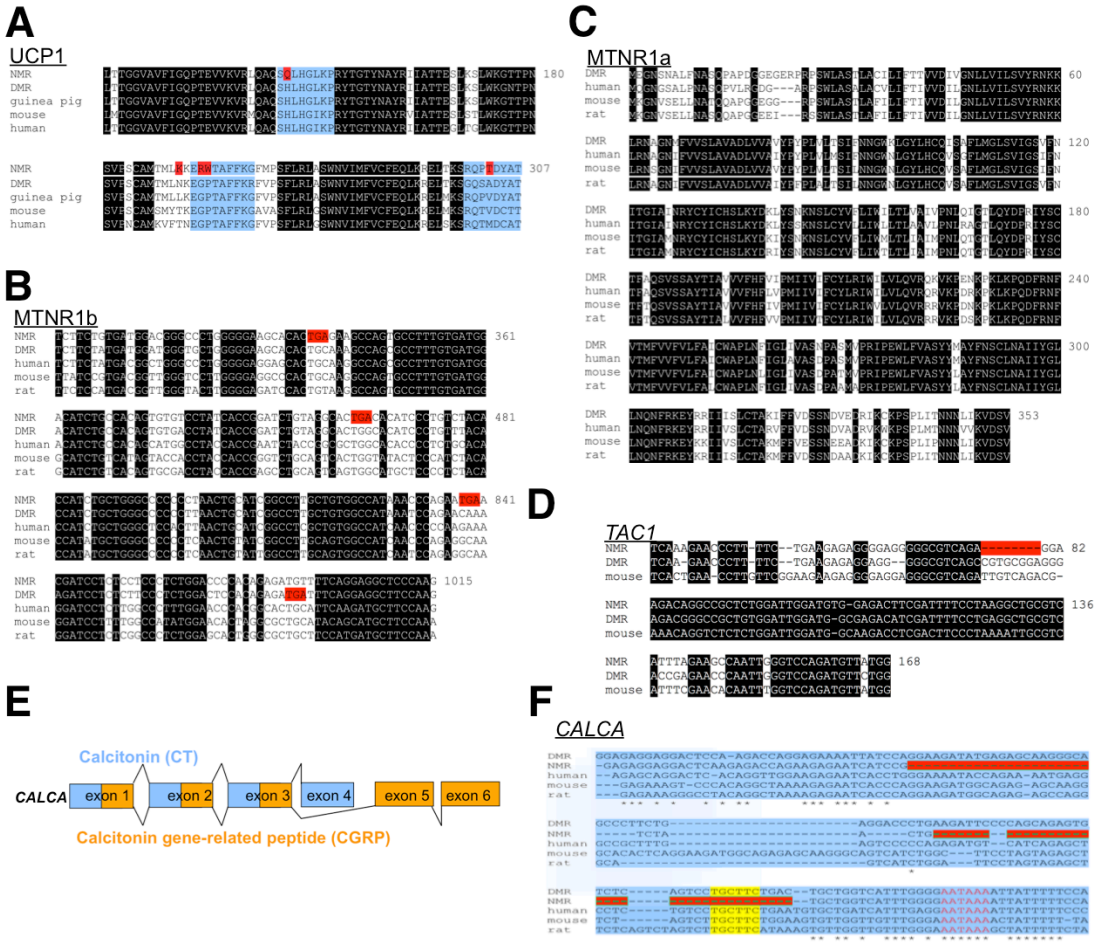
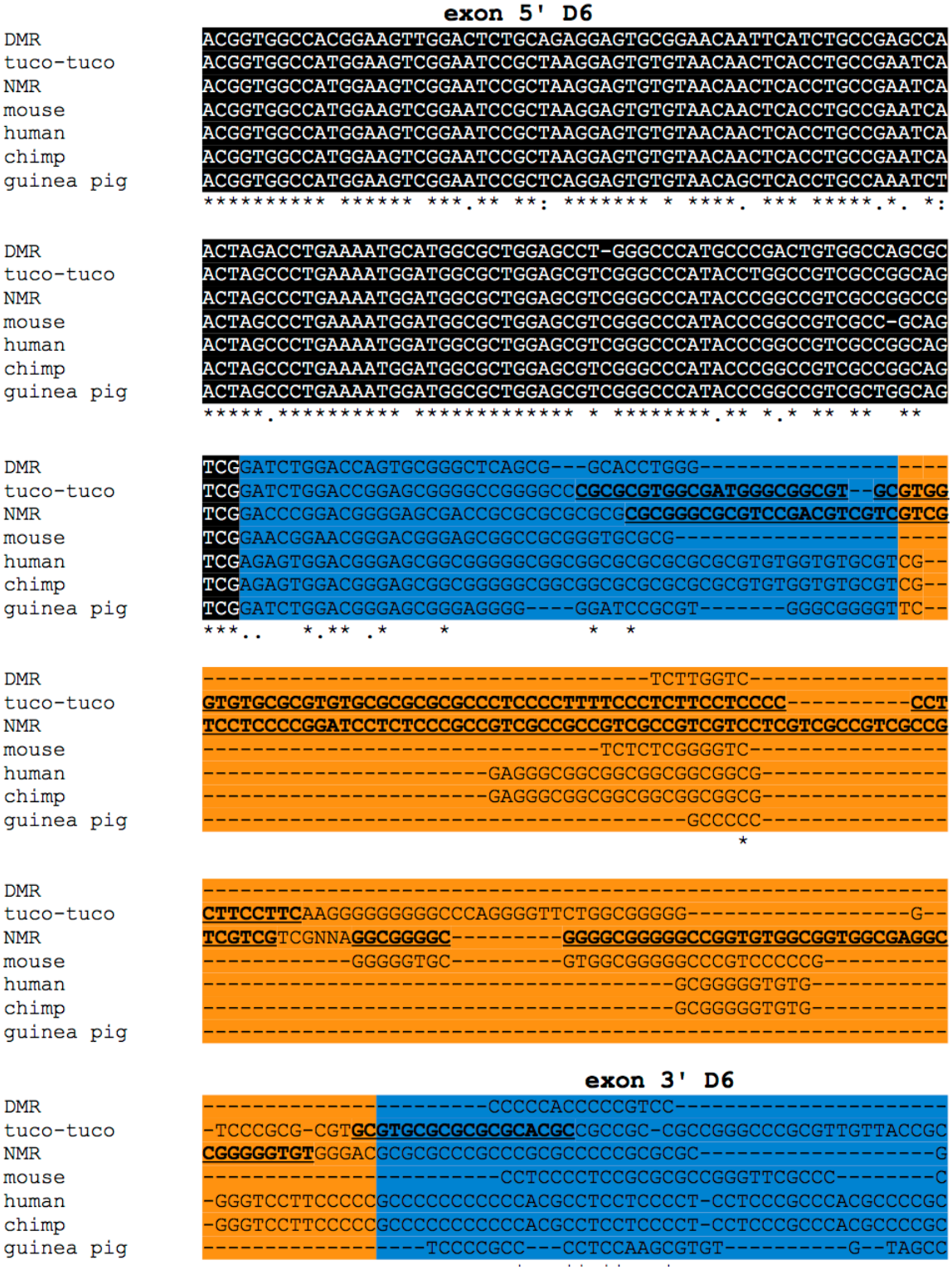


Figure S3, Related to Figure 4. Unique molecular changes in the naked mole rat

(A) Amino acid changes in functionally important motifs of the thermogenesis protein UCP1 are restricted to the NMR. Sequence alignment of rodent and human UCP1 is shown. Identical residues are shaded in black and conserved motifs are shown in blue. Amino acid residues unique to the poikilothermic NMR are highlighted in red. **(B)** Melatonin receptor 1b is inactivated in the DMR and NMR. Sequence alignment of rodent and human MTNR1b nucleotide sequences is shown. Identical residues are shaded in black, and premature stop codons are highlighted in red. **(C)** Melatonin receptor 1a is intact in the DMR. Sequence alignment of DMR and other mammalian MTNR1a sequences is shown. Identical residues are shaded in black. **(D)** A deletion within the promoter of *TAC1*, which encodes substance P (SP) is unique to the NMR and is not present in the related African mole rat, the DMR. Identical residues are shaded in black and the region deleted in NMR is highlighted in red. **(E)** Schematic genomic organization of human *CALCA* and the derived calcitonin/CT and calcitonin gene-related peptide/CGRP exons. Exons are shown as boxes. Exons employed by CT and GCRP are shown in blue and orange, respectively, while exons 1-3 shared by both isoforms are shown in both colors. The splicing pattern and the peptide sequences of prepro-calcitonin and prepro-calcitonin gene-related peptide are shown at the top and bottom of the exons, respectively. **(F)** NMR-specific deletions within exon 4 of *CALCA*. Multiple sequence alignment of NMR, DMR, mouse, rat and human exon 4 is shown. Exons are highlighted in blue and polyadenylation sites are shown in red. Regions deleted in the NMR are highlighted in red. Stars indicate nucleotide positions that are shared among all sequences. A highly conserved region deleted in the NMR is shown in yellow.


exon 5' D6
DMR
tuco-tuco
NMR
mouse
human
chimp
guinea pig
exon 3' D6

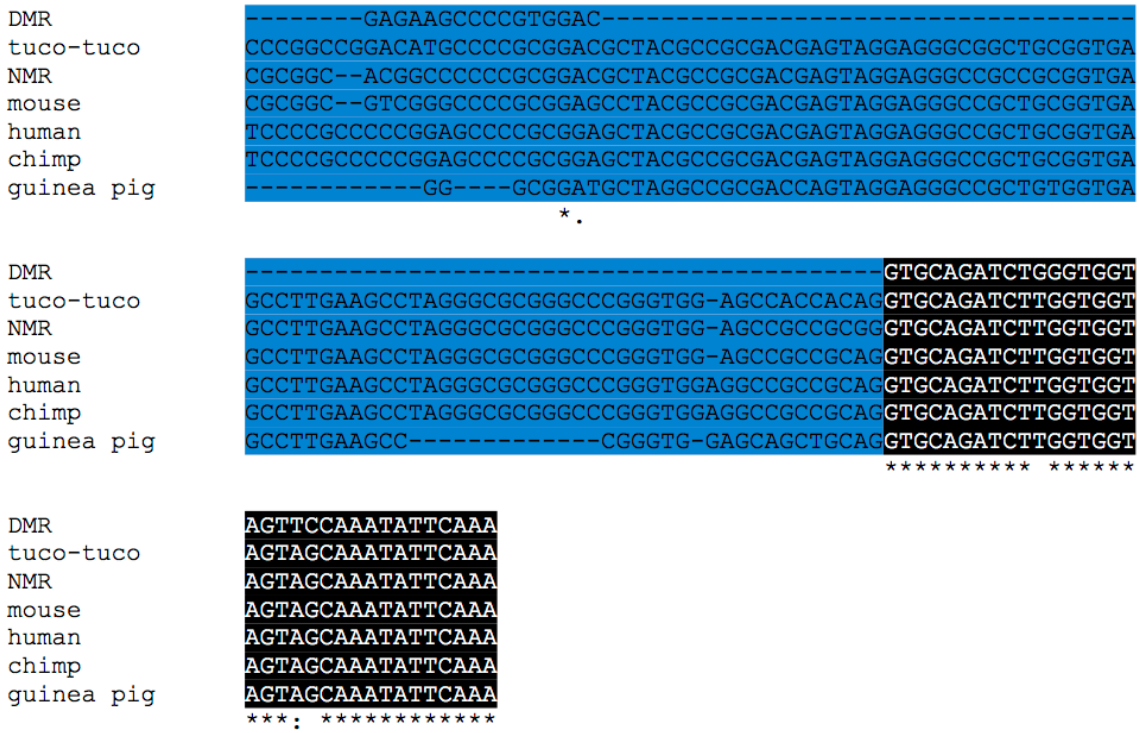


Figure S4, Related to Figure 5. Multiple sequence alignment of the mammalian D6 region of 28S ribosomal RNA

A highly conserved region of the mammalian 28S rRNA D6 domain is shown in black. Divergent regions of the D6 domain, exon 5' D6 and exon 3' D6, are shown in blue, and the cryptic intron is shown in orange. Dashes (–) indicate gaps, stars (*) indicate complete conservation and dots (.) indicate similarity. Repeat elements are indicated in bold and underlined.

Table S1, Related to Figure 1. Significantly expanded and contracted gene families in African mole rat lineages

(provided in a separate file).

Table S2, Related to Figure 1. Pseudogenes and lost genes in DMR, and pseudogenes common to the DMR and NMR

(provided in a separate file).

Table S3, Related to Figure 4. Genes under positive selection in the ancestral lineage of DMR/NMR, the DMR and the NMR, and associated GO categories

(provided in a separate file).

Table S4, Related to Figures 2 and 3. RPKM values and GO enrichment of genes differentially expressed in the liver of subterranean rodents

(provided in a separate file).

Table S5, Related to Figure 2. RPKM values and GO enrichment of genes differentially expressed in the brain of subterranean rodents

(provided in a separate file).

Table S6, Related to Figures 1 and 4. Lineage-specific accelerated GO categories of DMR and NMR ($P \leq 0.05$)

(provided in a separate file).

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

DMR and NMR genome features and evolution

Prediction of protein-coding genes

To predict genes in the DMR genome, we used both homology-based and *de novo* methods, and in addition, RNA-seq data were incorporated. For the homology-based prediction, human, mouse and rat proteins were downloaded from Ensembl (release 64) and mapped onto the genome using tBLASTn (Kent, 2002). Then, homologous genome sequences were aligned against the matching proteins using GeneWise (Birney et al., 2004) to define gene models. For *de novo* prediction, Augustus (Stanke, 2003) and GENSCAN (Salamov and Solovyev, 2000) were employed to predict coding genes, using appropriate parameters. RNA-seq data were mapped to genome using TopHat (Trapnell et al., 2009), and transcriptome-based gene structures were obtained by cufflinks (http://cufflinks.cbcb.umd.edu). Finally, homology-based, *de novo* derived and transcript gene sets were merged to form a comprehensive and non-redundant reference gene set using GLEAN (http://sourceforge.net/projects/glean-gene), removing all genes with sequences less than 50 amino acid as well as those that only had the *de novo* support. We obtained a reference gene set that contained 22,179 DMR genes, and the parameter of DMR genes is comparable to other published mammals, including the related naked mole rat (Kim et al., 2011) (Table 1).

Functional annotation of genes

Gene functions were assigned according to the best match of the alignments using BLASTp against SWISS-PROT databases (Bairoch and Apweiler, 1997). Gene Ontology IDs for each gene were obtained from the corresponding InterPro entry (Ashburner et al., 2000; Mulder and Apweiler, 2007). All genes were aligned against KEGG proteins, and the pathway in which the gene might be involved was derived from the matching genes in KEGG (Kanehisa and Goto, 2000).

Phylogenetic analysis

A gene family is a group of similar genes descended from a single gene in the last common ancestor of the targeted species. We used the TreeFam methodology (Li et al., 2006) to define gene families in 10 mammalian genomes (human, rhesus macaque, rabbit, mouse, rat, Chinese hamster, dog, guinea pig, DMR and NMR). We carried out the same pipeline and parameters that we used in our previously published work (Kim et al., 2011; Li et al., 2010a). In total, we obtained 19,839 gene families of which 6,133 are single-copy orthologous families. CDS sequences from each single-copy family were aligned guided by MUSCLE (Edgar, 2004) alignments of protein sequences and concatenated to one super gene for each species. Next, RAxML (Stamatakis et al., 2005) was applied to build phylogenetic trees under JTT+gamma model. 1,000 bootstrap replicates were employed to assess the branch reliability in RAxML. Divergence time was estimated by PAML MCMCTree (Yang and Rannala, 2006) using 4d site sequences of 5,838 single copy genes (Figure S1D).

Gene family expansion and contraction

To examine expansion and contraction of gene families in the DMR, we inferred the rate and direction of change of gene family size for DMR and a group of other mammals (human, rhesus, dog, NMR, rabbit, Chinese hamster, guinea pig, mouse and rat) using CAFÉ (De Bie et al., 2006), which is based on a stochastic birth-death model.

DMR gained and lost genes

To determine the orthologous relationship between DMR and human proteins, sequences of the human protein dataset were downloaded from Ensembl (release 64). The longest transcript was chosen to represent each gene with alternative splicing variants. We then subjected human and DMR proteins to BLASTp analysis with the similarity cutoff threshold of E-value=1e-5. With the human protein set as a reference, we found the best hit for each DMR protein, with the criteria that more than 30% of the aligned sequence showed an identity above 30%. Reciprocal best-match pairs were defined as orthologs. Then, gene order information was used to filter out false positive orthologs caused by incorrect draft genome assembly and annotation. Orthologs not in gene synteny blocks were removed from further analysis. For example, in the case of 3 continuous genes in the human genome A, B and C, even if all three orthologs could be identified in both humans and DMR based on the cutoff threshold described above, but the B gene in the DMR genome was not located between A and C genes, it was filtered out, as it could be located in another scaffold or another place within the same scaffold. Using this method, we identified gene synteny relationships for human and DMR lineages.

Orthology information was obtained as described above. Since it showed synteny information at the protein level, it could be used to analyze gene-gain and -loss between human and DMR lineages. In the protein synteny blocks, if a human protein had no DMR ortholog, and excluding false positive predictions that could be caused by annotation or genome assembly (gap $\geq$ 5%), this protein could be defined as either being lost in the DMR lineage or gained in the human lineage. Using DMR as a reference to generate the orthology relationship, we applied this procedure to identify the genes gained in the DMR lineage compared to the human lineage.

DMR pseudogenes

To detect homozygous pseudogenes in the DMR genome *in silico*, we first aligned all human genes (protein sequences downloaded from Ensembl (release 64) onto the DMR genome using BLASTp (with parameters -F F -e 1e-5). SOLAR (Sequential Oligogenic Linkage Analysis Routines) (Almasy and Blangero, 1998; Yu et al., 2006) was used to conjoin the fragmental alignments for each gene. Best hit regions of each gene with 5 kb flanking sequence were cut down and re-aligned with their corresponding human orthologous protein sequences using GeneWise (with the parameters "-genesf -tfor –quiet"), which helped define the detailed exon-intron structure of each gene (Birney et al., 2004). Genes with frameshifts and premature stop codons, as reported by GeneWise, were considered candidate pseudogenes. We further carried out a series of filtering processes: (i) To avoid the frameshifts and premature stop codons incorrectly reported due to flaws in the GeneWise algorithm, we also aligned all human proteins to their corresponding loci in the human genome using GeneWise. Genes with frameshifts and premature stop codons in the human-to-human alignment were filtered out; (ii) Using the results of the human-to-human alignment from GeneWise, candidate pseudogenes with obvious splicing errors near their frameshifts and premature stop codons were filtered out; (iii) Candidate pseudogenes with a low number of reads covering their frameshift and premature stop codon sites were considered assembly errors. In addition, cases with a considerable number of reads resulting from different genotypes at these sites were treated as heterozygous. The cases of assembly error and heterozygosis were filtered out.

Assessment for visual perception pseudogenes

To identify visual perception genes that were lost or pseudogenised in the DMR genome, we first employed GeneWise (Birney et al., 2004). To examine the visual perception pseudogenes, we estimated the rate ratio (ω) of non-synonymous to synonymous substitutions (Yang, 2007). The coding sequences of DMR visual perception genes were aligned with their human, mouse, rat and dog homologs using MUSCLE (Edgar, 2004). We used two branch models: Model H0, which considers all branches with the same ω, was used to compute the average ω; Model H1, which examined the DMR and other branches with different ω, was used to compute the ω of the DMR branch (ω2) and other branches (ω1). Then, the likelihood ratio test (LRT) was used to compute the p-value that rejected the model H0. If the p-value was less than 0.05, we could reject the model H0 and accept model H1.

Among the 13 identified visual perception genes, two were absent in the rat (*RP1L1* and *GRK7*), but other genes were present in all examined mammals (human, mouse, rat and dog). Visual genes inactivated or lost in the DMR are indicated in Table S2).

Identification of proteins with unique amino acid changes in African mole rats

To associate predicted DMR and NMR peptide sequences with human RefSeq IDs, as presented in UCSC multiway alignments, NCBI tBLASTn v2.2.26+ of the BLAST+ suite (Sayers et al., 2012) with an E-value cut-off set at 1e-5 was employed. The best match (≥50% overall amino acid sequence identity along the entire sequence and spanning ≥75% of the length of the query sequence) was used to annotate the sequences. African mole rat proteins were aligned to orthologs from 38 vertebrate species using ClustalW v2.1 (Larkin et al., 2007). In addition to the Damaraland mole rat (*Fukomys damarensis*) and the and the naked mole rat (*Heterocephalus glaber*), the following organisms obtained via the UCSC multiway track (Miller et al., 2007) were examined (the full common name includes the text in brackets): the little brown bat (*Myotis lucifugus*), the megabat/fruit bat/flying fox *Pteropus vampyrus*, human (*Homo sapiens)*, (common) chimpanzee (*Pan troglodytes*), (Sumatran) orangutan (*Pongo pygmaeus abelii*), rhesus monkey (*Macaca mulatta*), (hamadryas) baboon (*Papio hamadryas*), (common) marmoset (*Callithrix jacchus*), (Philippine) tarsier (*Tarsier syrichta*), (gray) mouse lemur (*Microcebus murinus*), bushbaby/small-eared greater galago (*Otolemur garnettii*), (Northern) tree shrew (*Tupaia belangeri*), (house) mouse (*Mus musculus*), (brown) rat (*Rattus norvegicus*), (Ord's) kangaroo rat (*Dipodomys ordii*), (domestic) guinea pig (*Cavia porcellus*), (thirteen-lined ground) squirrel (*Spermophilus tridecemlineatus*), (European) rabbit (*Oryctolagus cuniculus*), (American) pika (*Ochotona princeps*), alpaca (*Vicugna pacos*), cow (*Bos taurus*), horse (*Equus caballus*), cat (*Felis catus*), dog (*Canis lupus familiaris*), (European) hedgehog (*Erinaceus europaeus*), (common) shrew (*Sorex araneus*), (African) elephant (*Loxodonta africana*), (lesser hedgehog) tenrec (*Echinops telfairi*), (tammar) wallaby (*Macropus eugenii*), (gray short-tailed) opossum (*Monodelphis domestica*), platypus (*Ornithorhynchus anatinus*), (Western clawed) frog (*Xenopus tropicalis),* medaka/Japanese rice fish (*Oryzias latipes*), (three-spined) stickleback (*Gasterosteus aculeatus*), (tora)fugu (*Takifugu rubripes*), tetraodon/green spotted puffer (*Tetraodon nigroviridis*), zebrafish (*Danio rerio*), and sea lamprey (*Petromyzon marinus*).

In-house Perl scripts were used to parse the ClustalW output and identify unique amino acids. The Perl script scanned orthologous proteins for sites where types/groups of residues are shared by African mole rats only. The script groups residues into four groups: acidic (ED), basic (KHR), cysteine (C) and "other" (STYNQGAVLIFPMW). For example, it identifies cases where the NMR and DMR harbor E or D, whilst the other organisms contain basic, cysteine, or "other" residues at that particular site. Candidate proteins were manually examined for alignment quality and conservation of the region with amino acid changes.

The false positive rate of the detection of unique amino acids was estimated as follows: the error rate for SOAPdenovo assembly is assumed to be 1 nucleotide per 81,025 base pairs (Li et al., 2010b). Since only coding regions were used for amino acid conservation analysis, and the total predicted size of coding sequences (CDS) of the DMR and NMR is approximately 33 Mb, then 418 nucleotides could be considered as candidates for false positives. Potential tBLASTn misalignments are of minor importance since the parameters used were quite strict (E-value 1e-5). The unique amino acid method calls for a particular amino acid to be conserved among the tested vertebrate genomes, but differ in the genome-in-question (e.g. the DMR). In a test dataset, 3,287 amino acids were conserved among 36 genomes, out of 1,018,988 tested cases; therefore, only 0.32% of all sequences fit our search criteria. Taken together, with 418 candidates, one would expect a false positive rate of 1.33 per analysis; or in other words, one false positive per 315 unique amino acid residue candidates.

To manually validate the results pertaining to specific genes, we acquired genomic or transcriptome data of two additional African and two South-American hystricognath rodents: we generated brain RNA-seq data from the African mole rats Ansell's mole rat (*Fukomys anselli*, FA) and Mashona mole rat (*Fukomys darlingi*, FD). For comparison, we obtained the recently released genome sequence of the semi-subterranean degu (*Octodon degus*) (GenBank accession code AJSA00000000.1), as well as brain transcriptome data from the subterranean coruro (*Spalacopus cyanus*) of South-America. We acknowledge the efforts of the Broad Institute in assembling the genome sequence of the degu.

Orthologous gene set for evolutionary analysis using PAML

The orthologous sets of the 10 species from the phylogenetic tree were collected. For genes with alternative splice variants, the longest transcript was selected to represent the gene. Orthologs were identified using the method described in our previous study (Kim et al., 2011). MUSCLE (Edgar, 2004) was used for multi-protein sequence alignments of orthologs, using the human proteins as the reference. A series of filtering steps were performed on the alignments to assess alignment quality and conservation of exon-intron structure. Briefly, an alignment with more than 20% gaps or less than 50% identity between each protein and its human ortholog was identified as false positive and discarded. Next, the alignment was scanned exon-by-exon, and genes with more than 10% gaps or less than 60% identity, were manually checked against the genome sequences to validate whether these polymorphisms were real or caused by annotation errors. After these filtering processes, high quality alignments of protein sequences were retained and their alignments of associated codon sequences were used for further analysis. In total 9,367 1:1 ortholog alignments were obtained.

Identification of gene categories under accelerated evolution

The orthologous sets of the 10 mammals from the phylogenetic tree were used to identify proteins in each lineage that show accelerated evolution using a branch model implemented in the PAML (Phylogenetic Analysis by Maximum Likelihood) program. We aligned GO categories containing no fewer than 10 genes in the 10-species ortholog data set. PAML's free-ratio model (Yang, 2007) was used to calculate different ω ratios (also known as Ka/Ks) across every branch in the phylogenetic tree. To investigate whether fast-evolving genes in the NMR or DMR lineage were enriched for specific biological processes, a binomial test ($P \leq 0.05$) was used to identify GO categories with putatively accelerated non-synonymous divergence (increased dN/dS ratios) in either the DMR lineage or the NMR lineage.

Identification of genes under positive selection

To detect genes under positive selection in either the NMR or the DMR lineage, a series of evolutionary models were tested using PAML's branch model (Zhao et al., 2010). Briefly, the average ω across the tree (ω0), ω of the branch tested (ω2) and ω of all other branches (ω1) were determined with the following parameter settings: Codonfreq=2, kappa=2.5, initial omega=0.2. A chi-square test were used to determine if ω2 was significantly higher than ω1 and ω0, which implies that these genes may be fast-changed or fast-evolved in the appointed branch. After obtaining potential positively selected genes ($\omega 2 > \omega 1$ and p value ≤ 0.05 adjusted by the FDR method for multiple testing) (Benjamini and Yekutieli, 2001), we performed a final manual check of the alignment to confirm our results.

Expression analysis

Survey of gene expression in the brain

To identify genes that could play a role in the adaptation to an underground environment, we compared the brain transcriptomes of five subterranean and three surface-dwelling ('aboveground') hystricognath rodents. Total RNA was isolated from brain tissue (frontal lobe region) of *Fukomys damarensis* (DMR, n=2), *Fukomys anselli* (FA, n=1), *Fukomys darlingi* (FD, n=1), and *Spalacopus cyanus* (coruro, n=1). All animals were from captive colonies and the experiments were approved by the respective Animal Care and Use Committees (the University of Illinois at Chicago in the case of DMR; the University of South Bohemia in the case of FA, FD and coruro).

RNA sequencing libraries were constructed using the Illumina mRNA-Seq Prep Kit as per the manufacturer's instructions. Briefly, oligo(dT) magnetic beads were used to purify mRNA molecules. Next, mRNA was fragmented and randomly primed during the first strand synthesis by reverse transcription. Double-stranded cDNA fragments were then obtained by second-strand synthesis with DNA polymerase I, and the double stranded cDNA was subjected to end-repair by Klenow and T4 DNA polymerases, and finally A-tailed by Klenow lacking exonuclease activity. Ligation to Illumina Paired-End Sequencing adapters, size selection by gel electrophoresis and then PCR amplification completed the library preparation procedure. 200 bp paired-end libraries were sequenced using Illumina HiSeq 2000 (90 bp at each end).

Because no reference genomes are available for FA, FD and the coruro, the brain transcriptomes were assembled *de novo*. We also assembled paired-end reads from published RNA-seq data from the naked mole rat *Heterocephalus glaber* (NMR, n=2) (Kim et al., 2011) (NCBI GEO accession no. GSE30337) and the South-American

hystricognath rodents the domestic guinea pig *Cavia porcellus* (n=3) and the Brazilian guinea pig *Cavia aperea* (n=3), as well as a tame line of the brown rat (*Rattus norvegicus*; n=3) in the rodent family Muridae (Albert et al., 2012) (ArrayExpress accession no. E-MTAB-1249). With the exception of the NMR, all samples examined were males. Since many genes would be expected to change their expression between the surface-dwelling guinea pig (Rodentia, Hystricognathi) and the rat (Rodentia, Muridae) (Konopka et al., 2012), the present study included the rat as an outgroup in order to further narrow down the candidate driver genes for adaptation to a subterranean environment.

Reads were preprocessed to remove adapters and overrepresented sequences identified by FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). The software package Trimmomatic (Lohse et al., 2012) was used to trim Illumina reads falling below the established quality threshold (default settings) as well as to trim adapter sequences and remove unpaired reads. Paired reads were concatenated, keeping only reads with ≥36 bp length (after merging and adapter trimming). Next, transcriptome assembly was completed using the short-read *de novo* assembler Trinity (Grabherr et al., 2011). Abundance estimates were obtained by running RSEM (RNA-seq by Expectation Maximization) separately for the left and right reads from each paired-end run (sample) (Haas et al., 2013; Li and Dewey, 2011). A database of mouse UniRef90 protein sequences (Wu et al., 2006) was interrogated using transcripts predicted by Trinity as a query (BLASTx E-value cut-off 1e-20) (Grabherr et al., 2011). Next, The RSEM output files (RSEM.genes.results) were parsed using a custom Perl script, retaining Trinity transcripts that aligned across at least 80% of the corresponding mouse protein sequence (Grabherr et al., 2011). We did not exclude any genes from the annotations during quantification (e.g. paralogs were kept), and did not modify the gene annotations or attempt to identify new genes using our data.

Differentially expressed genes were detected using the method described by Chen and colleagues (Chen et al., 2010), which is based on the Poisson distribution (Audic and Claverie, 1997) and normalization for differences in the RNA output size and sequencing depth between samples, as well as accounting for different gene length. Genes with at least a two-fold difference in expression between the subterranean rodents (DMR, FD, FD, NMR and coruro) compared to the mouse and guinea pigs, with a false discovery rate (FDR) ≤ 0.05, were defined as differentially expressed genes. The resulting gene set was manually curated to remove likely false positive calls (e.g., if FA has a very low RPKM value compared the other subterranean rodents and is similar to guinea pigs and the rat). While we appreciate that, unlike techniques such as *in situ* hybridization and laser microdissection-coupled RNA-seq, transcriptome analysis of a tissue pieces from a complex organ such as the brain may not correspond to a uniform cell population, we believe that this work provides a foundation for the study of genes that that could play a role in the adaptations to a subterranean environment.

Survey of globin gene expression response to short-term hypoxia

DMR, NMR and rat (n=3) were exposed to short-term hypoxia (8 hours) at 8% O_2, the estimated oxygen condition in NMR tunnels (Bennet and Faulkes, 2000). Samples were sequenced, processed and analyzed exactly as detailed above. *HIF3A*, which displays rapid elevated systemic mRNA expression upon short-term hypoxia in rats (Heidbrederet al., 2003), was used to confirm a hypoxia response.

Survey of gene expression in the liver

Total RNA was isolated from the liver from 3- to 7-year-old female and male DMRs. Transcriptome sequencing was performed as outlined above. Briefly, in addition to the DMR, liver RNA-seq data generated from three additional species were obtained: NRM (Kim et al., 2011) (NCBI GEO accession no. GSE30337), and the mouse and rat (Merkin et al., 2012) (NCBI GEO accession no. GSE41637). The RNA-seq data analyzed comprised at least three individuals from each species. Transcriptome reads were mapped to reference genomes using TopHat (Trapnell et al., 2009), and mapped reads were analyzed using in-house Perl scripts. We identified orthologs and obtained expression data using gene annotations spanning coding regions. Differentially expressed genes were detected as outlined above. Genes with at least a two-fold difference in expression between the NMR and DMR compared to the mouse and rat, with a false discovery rate (FDR) ≤ 0.05, were defined as differentially expressed genes.

Gene enrichment analysis

Gene Ontology (GO) term analyses were performed using DAVID (Database for Annotation, Visualization and Integrated Discovery) (Huang da et al., 2009). Briefly, to test for enrichment we compared genes that were significantly differentially expressed in the subterranean rodents against DAVID's GO FAT database. The DAVID functional annotation tool categorizes GO terms and calculates an "enrichment score' or EASE score (a modified Fisher's exact test-derived p-value). Categories with smaller p-values ($P < 0.01$) (Kosiol et al., 2008; Qiu et al., 2012) and larger fold-enrichments (≥ 2.0) were considered interesting and most likely to convey biological meaning (Huang da et al., 2009).

Western blot analysis for haemoglobin α in rodent brain tissue

NMR, mouse, rat and guinea pig brains (frontal lobes) were used for preparation of protein lysate. Tissues were rinsed in cold PBS to minimize blood contamination, homogenized in RIPA buffer (Abcam) with protease inhibitor (Sigma), centrifuged 10 min at 150,000 rpm and total protein concentration in soluble fraction was determined by Coomassie Protein Assay Reagent (Thermo Scientific). 50 μg of total protein from each tissue samples was resolved by SDS-PAGE on 10% Bis-Tris gel in non-reducing conditions and transferred onto a PVDF membrane (Invitrogen). Goat polyclonal antibody against mouse hemoglobin α-chain (sc-31333, Santa-Cruz) was used as a primary antibody in dilution 1:500 and donkey anti-goat IgG-HRP (horseradish peroxidase) (sc-2020, Santa-Cruz) was used as a secondary in dilution 1:1000. Monoclonal anti-β-actin antibody (A2228, Sigma) was used for loading control. Western blots were developed with Clarity Western ECL blot substrate (Bio-Rad) and visualized with ChemiDoc XRS+ imaging system (Bio-Rad).

Identification of the novel 28S RNA in the NMR

Denaturing agarose gel electrophoresis

Total RNA from the NMR, DMR and mouse was isolated with RNAqueous Midi total RNA isolation kit (Invitrogen) using 50-100 mg of frozen tissue. The resulting RNA samples (1 μg of total RNA) were electrophoresed on 1% denaturing agarose gels and stained with ethidium bromide. RNA Millennium Marker (Invitrogen) was used as a molecular weight marker.

Characterization of the 5` end of the 28S NMR fragment

Total RNA from the NMR liver was extracted with TRIzol reagent (Invitrogen) according to the manufacturer`s protocol. A 5` RACE RNA adapter (5`-GCUGAUGGCGAUGAAUGAACACUGCGUUUGCUGGCUUUGAUGAAA-3`) was ligated with T4 RNA ligase 1 (NEB) in the following reaction mixture, which was incubated for 2 hours at 37 °C: 300 ng total RNA, 50 pmol 5` RACE adapter 1 μl PEG 8000, 1 μl buffer, 1 μl T4 RNA ligase 1 enzyme, and water to 10 μl. The reaction product was purified by phenol-chloroform method and precipitated by ethanol, and the resulting RNA was used for PCR using AccuPrime polymerase (Invitrogen) with primers 5RACEouter (5`-GCTGATGGCGATGAATGAACACTG-3`) and D6R (5`-GACTGACCCATGTTCAACTGCTGT-3`). Cycling conditions included 25 cycles at 60° for 30 sec, with elongation at 72° over 20 sec. PCR products were purified using a QIAGEN purification kit, subcloned into *pGEM-T Easy* (Promega) and transformed into DH5α *E.coli* strain (Invitrogen). Colonies were grown in LB media, and the plasmids purified with a QIAGEN miniprep kit and sequenced by capillary electrophoresis.

Characterization of the 3` end of the 28S NMR RNA fragment

300 ng of total RNA from the NMR liver was treated with polyA polymerase (NEB) with 1 mM ATP to attach a polyA tail to ribosomal RNA fragments. After ethanol precipitation, reverse transcription and second strand synthesis was set up using Invitrogen`s SuperScript Double-Stranded cDNA Synthesis kit. DNA was purified by phenol-chloroform and precipitated with ethanol. DNA ends were blunted with Klenow Polymerase (Fermentas) according to the manufacturer`s suggestions. After blunting, DNA fragments were phosphorylated by T4 kinase to enable subcloning into blunted *pGEM-T Eas*y vector (Promega). Plasmids were prepared for sequencing as described above.

Multiple sequence alignments

BLASTn (Altschul et al., 1990) was used to interrogate nucleotide databases for sequences similar to the NMR and Talas tuco-tuco. 28S rRNA sequences of the DMR, NMR, mouse (GenBank accession no. AL355742), human (GenBank accession no. M11167), chimpanzee (GenBank accession no. K03429), guinea pig (GenBank accession no. NT_176417), Talas tocu-tocu (GenBank accession no. AF119340 cDNA and AF119339 genomic) were aligned using Clustal Omega (Sievers et al., 2011). Repeat elements were identified using RepeatMasker (Tarailo-Graovac and Chen, 2009).

Protein modeling

For protein modeling of ARG1 (Figure 2B), we used Molsoft ICM Browser Pro software (http://www.molsoft.com/icm_browser_pro.html) and a crystal structure model of human arginase I (D'Antonio et al., 2011), available from the RCSB Protein Data Bank (http://www.pdb.org) (PDB entry 3tf3) (Berman et al., 2003).

Validation of genes by PCR

The following genes were selected for validation by PCR and Sanger sequencing of NMR genomic DNA or cDNA: *UCP1* (5`-CTCTTTCTGTTTGGCTCCTTGAGGGACG-3 and 5`-TGTATGCCAAATCCTTTTCTGGAAGCTGA-3; 981 bp; cDNA template), *CDKN2A* (5`-ATCCGAGACTTTTAAGGTTGTCTG-3 and 5`-

ATCCGAGACTTTTAAGGTTGTCTG-3; 834 bp; cDNA template), *MTNR1A* (5`-TGACAAGAATTCGCTCTGCT-3 and 5`-GACCTTGAGCAATACAGGTAACTATTAATAATACT-3; 675 bp; gDNA template), *MTNR1B* (5`- GTAATTTATCCTTGGTGAGTCTGGCATTG-3 and 5`-ATTCATAGGTTTGTTGCCGCTGTAGAGCAGA-3; 903 bp; gDNA template), *GJA10* (5`-AGACTAGAAGAGGGTAAGAAGTAAACTGAG-3 and 5`-GAGAGAATATTGGCCTATCATAAACCCTAC-3; 792 bp; gDNA template)

SUPPLEMENTAL REFERENCES

Albert, F.W., Somel, M., Carneiro, M., Aximu-Petri, A., Halbwax, M., Thalmann, O., Blanco-Aguiar, J.A., Plyusnina, I.Z., Trut, L., Villafuerte, R., et al. (2012). A comparison of brain gene expression levels in domesticated and wild animals. PLoS Genet. *8*, e1002962.

Almasy, L., and Blangero, J. (1998). Multipoint quantitative-trait linkage analysis in general pedigrees. Am. J. Hum. Genet. *62*, 1198-1211.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403-410.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T.*, et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genet. *25*, 25-29.

Audic, S., and Claverie, J.M. (1997). The significance of digital gene expression profiles. Genome Res. *7*, 986-995.

Bennett, N.C., and Faulkes, C.G. (2000). African mole-rats: ecology and eusociality (Cambridge, UK: Cambridge University Press).

Bairoch, A., and Apweiler, R. (1997). The SWISS-PROT protein sequence data bank and its supplement TrEMBL. Nucleic Acids Res. *25*, 31-36.

Benjamini, Y., and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. Annal. Stat. 29, 1165-1188.

Berman, H., Henrick, K., and Nakamura, H. (2003). Announcing the worldwide Protein Data Bank. Nature Struct. Biol. *10*, 980.

Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and Genomewise. Genome Res. *14*, 988-995.

Chen, S., Yang, P., Jiang, F., Wei, Y., Ma, Z., and Kang, L. (2010). De novo analysis of transcriptome dynamics in the migratory locust during the development of phase traits. PloS ONE *5*, e15633.

D'Antonio, E.L., and Christianson, D.W. (2011). Crystal structures of complexes with cobalt-reconstituted human arginase I. Biochemistry *50*, 8018-8027.

De Bie, T., Cristianini, N., Demuth, J.P., and Hahn, M.W. (2006). CAFE: a computational tool for the study of gene family evolution. Bioinformatics *22*, 1269-1271.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792-1797.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q.*, et al.* (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biot. *29*, 644-652.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M.*, et al.* (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Prot. *8*, 1494-1512.

Heidbreder M., Fröhlich F., Jöhren O., Dendorfer A., Qadri F., Dominiak P. (2003). Hypoxia rapidly activates HIF-3alpha mRNA expression. FASEB J. *11*, 1541-3.

Huang da, W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Prot. *4*, 44-57.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. *28*, 27-30.
Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. Genome Res. *12*, 656-664.
Kim, E.B., Fang, X., Fushan, A.A., Huang, Z., Lobanov, A.V., Han, L., Marino, S.M., Sun, X., Turanov, A.A., Yang, P.*, et al.* (2011). Genome sequencing reveals insights into physiology and longevity of the naked mole rat. Nature *479*, 223-227.
Konopka, G., Friedrich, T., Davis-Turak, J., Winden, K., Oldham, M.C., Gao, F., Chen, L., Wang, G.Z., Luo, R., Preuss, T.M.*, et al.* (2012). Human-specific transcriptional networks in the brain. Neuron *75*, 601-617.
Kosiol, C., Vinar, T., da Fonseca, R.R., Hubisz, M.J., Bustamante, C.D., Nielsen, R., and Siepel, A. (2008). Patterns of positive selection in six Mammalian genomes. PLoS Genet. *4*, e1000144.
Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R.*, et al.* (2007). Clustal W and Clustal X version 2.0. Bioinformatics *23*, 2947-2948.
Li, H., Coghlan, A., Ruan, J., Coin, L.J., Heriche, J.K., Osmotherly, L., Li, R., Liu, T., Zhang, Z., Bolund, L.*, et al.* (2006). TreeFam: a curated database of phylogenetic trees of animal gene families. Nucleic Acids Res. *34*, D572-580.
Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., Huang, Q., Cai, Q., Li, B., Bai, Y.*, et al.* (2010a). The sequence and de novo assembly of the giant panda genome. Nature *463*, 311-317.
Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K.*, et al.* (2010b). De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. *20*, 265-272.
Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323.
Lohse, M., Bolger, A.M., Nagel, A., Fernie, A.R., Lunn, J.E., Stitt, M., and Usadel, B. (2012). RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. Nucleic Acids Res. *40*, W622-627.
Merkin, J., Russell, C., Chen, P., and Burge, C.B. (2012). Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. Science *338*, 1593-1599.
Miller, W., Rosenbloom, K., Hardison, R.C., Hou, M., Taylor, J., Raney, B., Burhans, R., King, D.C., Baertsch, R., Blankenberg, D.*, et al.* (2007). 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. Genome research *17*, 1797-1808.
Mulder, N., and Apweiler, R. (2007). InterPro and InterProScan: tools for protein sequence classification and comparison. Methods Mol. Biol. *396*, 59-70.
Qiu, Q., Zhang, G., Ma, T., Qian, W., Wang, J., Ye, Z., Cao, C., Hu, Q., Kim, J., Larkin, D.M.*, et al.* (2012). The yak genome and adaptation to life at high altitude. Nature Genet. *44*, 946-949.
Salamov, A.A., and Solovyev, V.V. (2000). Ab initio gene finding in Drosophila genomic DNA. Genome Res. *10*, 516-522.
Sayers, E.W., Barrett, T., Benson, D.A., Bolton, E., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., Dicuccio, M., Federhen, S.*, et al.* (2012). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. *40*, D13-25.
Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J.*, et al.* (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. *7*, 539.

Stamatakis, A., Ludwig, T., and Meier, H. (2005). RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics *21*, 456-463.

Stanke, M., and Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics *19 Suppl 2*, ii215-225.

Tarailo-Graovac M., and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. In Current Protocols in Bioinformatics, A. D Baxevanis *et al.* ed. (Hoboken, NJ: John Wiley & Sons), pp. 4.10.1-4.10.14.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111.

Wu, C.H., Apweiler, R., Bairoch, A., Natale, D.A., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R.*, et al.* (2006). The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res. *34*, D187-191.

Yang, Z., and Rannala, B. (2006). Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. Mol. Biol. Evol. *23*, 212-226.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. *24*, 1586-1591.

Yu, X.J., Zheng, H.K., Wang, J., Wang, W., and Su, B. (2006). Detecting lineage-specific adaptive evolution of brain-expressed genes in human using rhesus macaque as outgroup. Genomics *88*, 745-751.

Zhao, H., Yang, J.R., Xu, H., and Zhang, J. (2010). Pseudogenization of the umami taste receptor gene Tas1r1 in the giant panda coincided with its dietary switch to bamboo. Mol. Biol. Evol. *27*, 2669-2673.