LETTER

doi:10.1038/nature10533

Genome sequencing reveals insights into physiology and longevity of the naked mole rat

Eun Bae Kim[1]*, Xiaodong Fang[2]*, Alexey A. Fushan[1]*, Zhiyong Huang[2]*, Alexei V. Lobanov[3], Lijuan Han[2], Stefano M. Marino[3], Xiaoqing Sun[2], Anton A. Turanov[3], Pengcheng Yang[2], Sun Hee Yim[3], Xiang Zhao[2], Marina V. Kasaikina[3], Nina Stoletzki[3], Chunfang Peng[2], Paz Polak[3], Zhiqiang Xiong[2], Adam Kiezun[3], Yabing Zhu[2], Yuanxin Chen[2], Gregory V. Kryukov[3,4], Qiang Zhang[2], Leonid Peshkin[5], Lan Yang[2], Roderick T. Bronson[6], Rochelle Buffenstein[7], Bo Wang[2], Changlei Han[2], Qiye Li[2], Li Chen[2], Wei Zhao[2], Shamil R. Sunyaev[3,4], Thomas J. Park[8], Guojie Zhang[2], Jun Wang[2,9,10] & Vadim N. Gladyshev[1,3,4]

Maido Remm

Bioinformaatika Journal Club

17.09.2012

Eesti keeles: paljastuhnur

The naked mole rat (*Heterocephalus glaber)* is strictly subterranean rodent, found in the dry, tropical grasslands that cover Kenya, Ethiopia and Somalia.

They **live in full darkness, at low oxygen and high carbon dioxide concentrations**, and are **unable to sustain thermogenesis** nor **feel certain types of pain**.

http://theinfinitevariety.wordpress.com/2010/11/14/it%E2%80%99s-all-in-the-family-the-naked-mole-rat/

The naked mole rat is an extraordinarily long-lived eusocial mammal. Although it is the size of a mouse, its **maximum lifespan exceeds 30 years**, making this animal the longest-living rodent. Naked mole rats show negligible senescence, **no age-related increase in mortality**, and high fecundity until death. In addition to delayed ageing, they are **resistant to both spontaneous cancer and experimentally induced tumorigenesis**.

Naked mole rats naturally **reside in large colonies** with a **single breeding female, the 'queen', who suppresses the sexual maturity of her subordinates**.

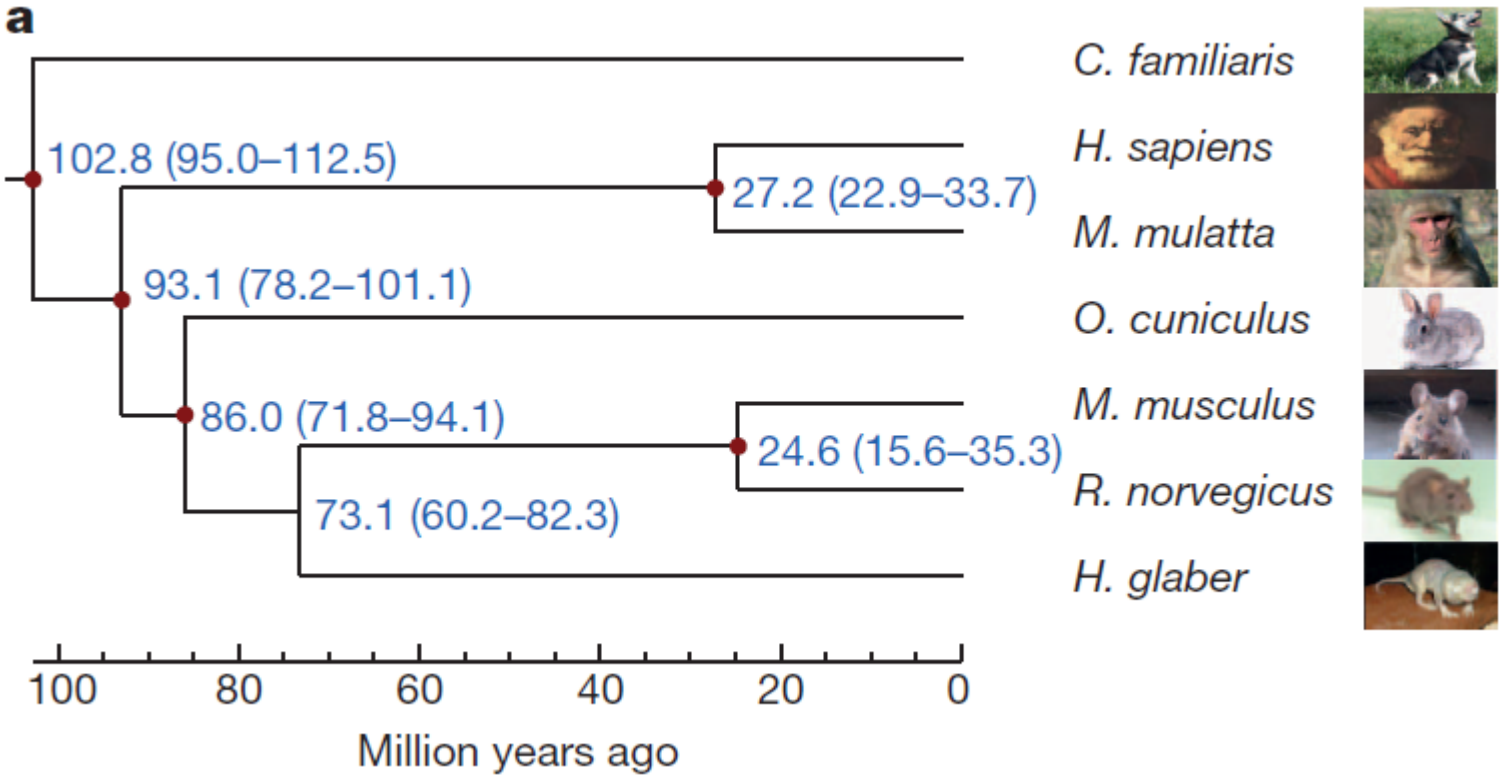


Figure 1. Relationship of the NMR to other mammals. Estimation of the time of divergence (with error range shown in parentheses).

The BRMC approach was used to estimate the species divergence time using the program MULTIDIVTIME, which was implemented using the Thornian Time Traveller (T3) package (ftp://abacus.gene.ucl.ac.uk/pub/T3/).

What type of sequence data was produced?

Genomic sequence:

ca 500 Gbp from single non-breeding male using Illumina HiSeq 2000 platform

Transcriptome sequences:

Tissue variation (brain, kidney, liver)

Age variation (newborn, 4-year old, 20-year old)

Oxygen variation (normal, 8% oxygen for 1 week)

Supplementary Table 20. Transcriptome sequencing data statistics.

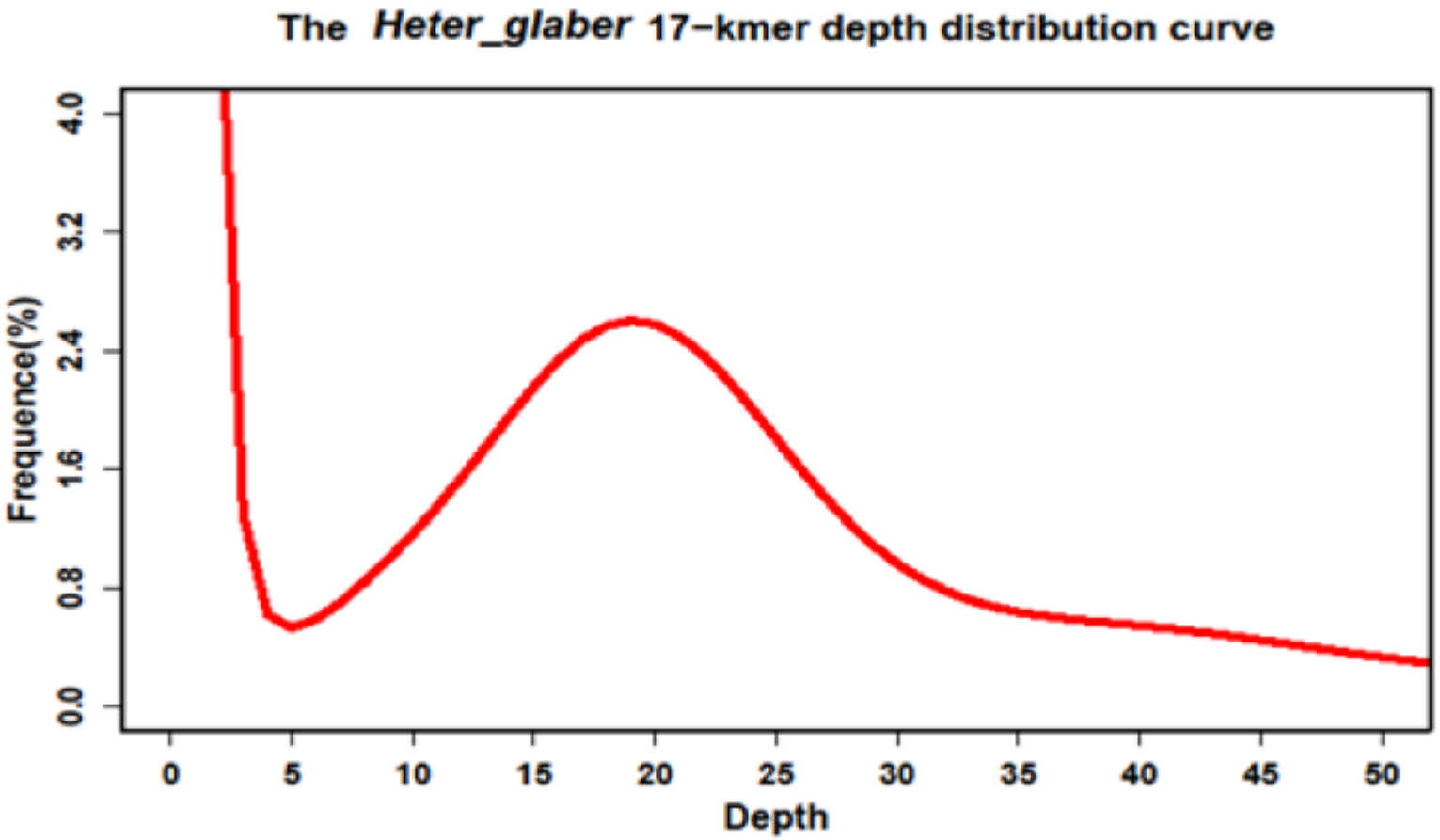
	Total reads (M)	Total base (G)	Map reads (M)	Reads (%)	Map base (G)	Base (%)	Genome coverage (%)
New-brain	55.1	4.96	47.8	86.8	4.06	81.9	3.96
New-kidney	48.2	4.34	42.5	88.2	3.63	83.6	4.38
New-liver	53.3	4.8	45.7	85.7	3.85	80.2	3.22
4-brain	53.4	4.81	43.7	81.8	3.64	75.7	3.19
4-kidney	50.4	4.54	40.5	80.4	3.35	74	2.91
4-liver	54.5	4.91	45.2	83	3.76	77	2.68
20-brain	58.4	5.25	48.2	82.5	4.05	77.1	3.89
20-liver	52.8	4.75	44.9	85	3.78	80	3.11
20-kidney	56	5	45.4	81.7	3.8	76	3.2
Low-liver	66.7	6	55.67	83.5	4.63	77.2	2.41
Low-kidney	65.8	5.93	52.09	79.1	4.33	73.1	3.4
Low-brain	63.8	5.74	51.9	81.4	4.36	75.9	3.61

New refers to a newborn NMR, 4 and 20 indicate the age of animals, and low indicates that samples were taken from an animal subjected to 8% O_2.

Table 1 | Global statistics of the NMR genome

Sequencing	Insert size (bp)	Total data (Gb)	Sequence coverage (fold)
Paired end	170–800	126.52	47
libraries	2–20 $\times 10^3$	120.66	45
	Total	247.18	92
Assembly	**N50 (kb)**	**Longest (kb)**	**Size (Gb)**
Contigs	19.3	179	2.45
Scaffolds	1,585	7,787	2.66
Annotation	**Number**	**Total length (Mb)**	**Percentage of the genome**
Repeats	3,090,116	666.7	25
Genes	22,561	722.3	27.1
CDS	181,641	32.5	1.2

In total, we generated about 475.78G of sequence, and following filtering out low quality and duplicated reads, 247G (90x coverage) was retained for assembly.



The genome size, G, was defined as G=K_num/K_depth, where the K_num is the total number of k-mers, and K_depth is the frequency occurring more frequently than other frequencies. In the present study, K is 17, K_num is 52,143,337,243 and K_depth is 19; thus, the NMR genome size is estimated to be 2.74G, which is comparable to that of other rodents.

1.3 Genome assembly

The NMR genome was assembled *de novo* using SOAPdenovo with k=41.

Low quality reads were filtered out and potential sequencing errors were removed or corrected by k-mer frequency methodology. We filtered out the following type of reads:

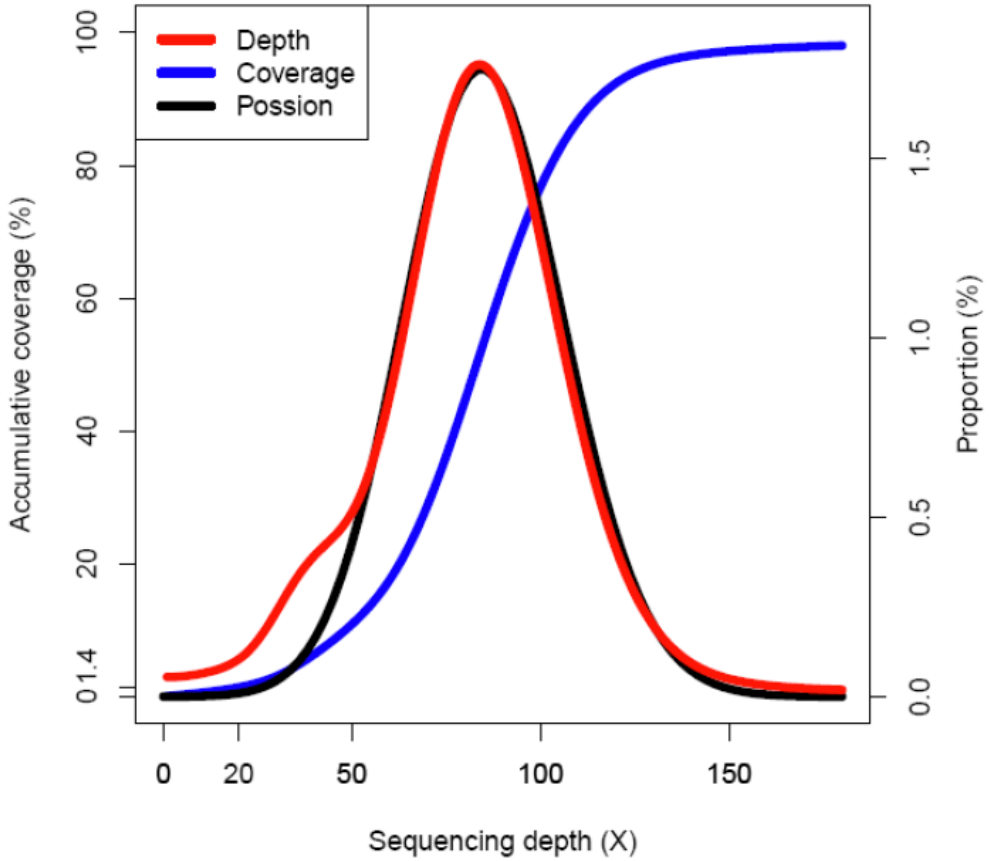
1. Reads having a 'N' over 10% of its length.
2. Reads from short insert-size libraries having more than 65% bases with the quality ≤ 7, and the reads from large insert-size libraries that contained more than 80% bases with the quality ≤ 7.
3. Reads with more than 10 bp from the adapter sequence (allowing no more than 2 bp mismatches).
4. Small insert size paired-end reads that overlapped ≥ 10 bp between the two ends.
5. Read 1 and read 2 of two paired-end reads that were completely identical (and thus considered to be the products of PCR duplication).
6. Reads having k-mer frequency <4 after correction (to minimize the influence of sequencing errors).

476 Gbp -> 247 Gbp

Genome assembly quality control

1. Completeness:
97.4% reads could be mapped back to the assembled genome.

2. Abnormalities of distribution (collapsed regions):
Distribution of coverage follows expected distribution (mode = 88x coverage).
Approximately 98.6% of the genome was covered by at least 20 reads**.**



Gene prediction methods

To predict genes in the NMR genome, we used both homology-based and *de novo* methods. In addition, RNA-seq data were incorporated. For the homology-based prediction, human and mouse proteins were downloaded from Ensembl (release 56) and mapped onto the genome using TblastN. Then, homologous genome sequences were aligned against the matching proteins using Genewise to define gene models. For *de novo* prediction, Augustus and Genscan were employed to predict coding genes, using appropriate parameters. RNA-seq data were mapped to genome using Tophat, 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 *de novo* support.**

Supplementary Table 6. Statistics of predicted protein-coding genes.

Species	*Gene set number*	*Complete ORF*	*%*	*Single exon gene*	*%*	*Average transcript length (bp)*	*Average ORF length (bp)*	*Average exons per gene*	*Average exon length (bp)*	*Average intron length (bp)*
NMR	22,561	19,137	84.82	3,930	17.42	32,533	1,439	8.05	178.73	4,410
Human	22,389	20,098	89.77	3,318	14.82	44,855	1,560	8.96	174.08	5,436
Mouse	23,317	21,196	90.9	4,648	19.93	33,684	1,481	8.37	176.82	4,366
Rat	22,841	16,745	73.31	3,552	15.55	30,892	1,452	8.59	169.06	3,879

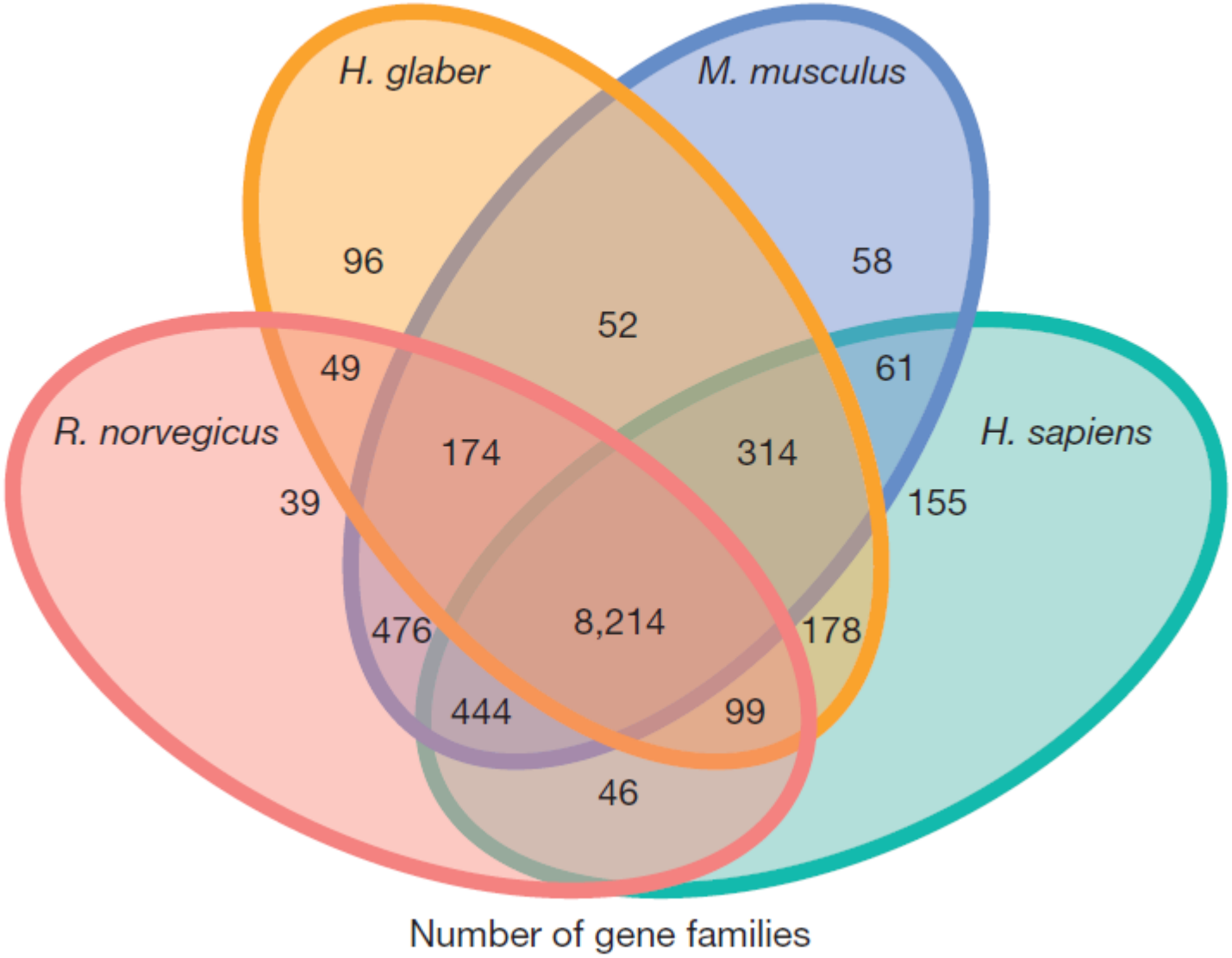


Figure 2. Common and unique NMR gene families.

Functional analysis methods:

- **missing and gained genes, pseudogenes**

(NMR wrt human genome)

- **Unique AA or DNA changes**

(unique AA or promoter DNA changes in positions where all mammals have conserved sequence)

- **positive selection regions**

(regions where nonsyn_subst_rate/syn_subst_rate >> 1)

- **mRNA expression analysis**

(old and young age; low and normal oxygen)

Functional analysis methods: missing and gained genes

Analysis of syntenic regions identified 750 gained and 320 lost NMR genes. We also identified 244 pseudogenes, containing frameshift and premature termination events.

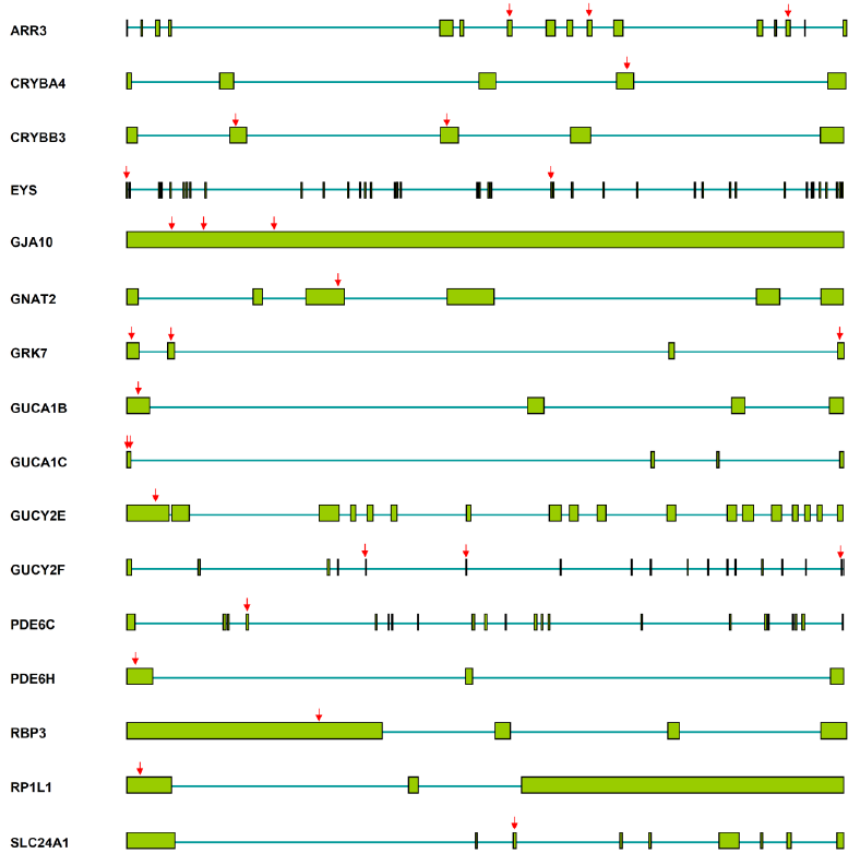
GO enrichment of genes that were lost in NMR.

GO_ID	GO_Term	GO_Class	Adjusted p-value
GO:0030529	ribonucleoprotein complex	CC	0.023655
GO:0003735	structural constituent of ribosome	MF	0.023655
GO:0005840	ribosome	CC	0.023655
GO:0004550	nucleoside diphosphate kinase activity	MF	0.023655
GO:0006183	GTP biosynthetic process	BP	0.023655
GO:0006228	UTP biosynthetic process	BP	0.023655
GO:0006241	CTP biosynthetic process	BP	0.023655
GO:0006412	translation	BP	0.046916

GO enrichment of pseudogenes in NMR.

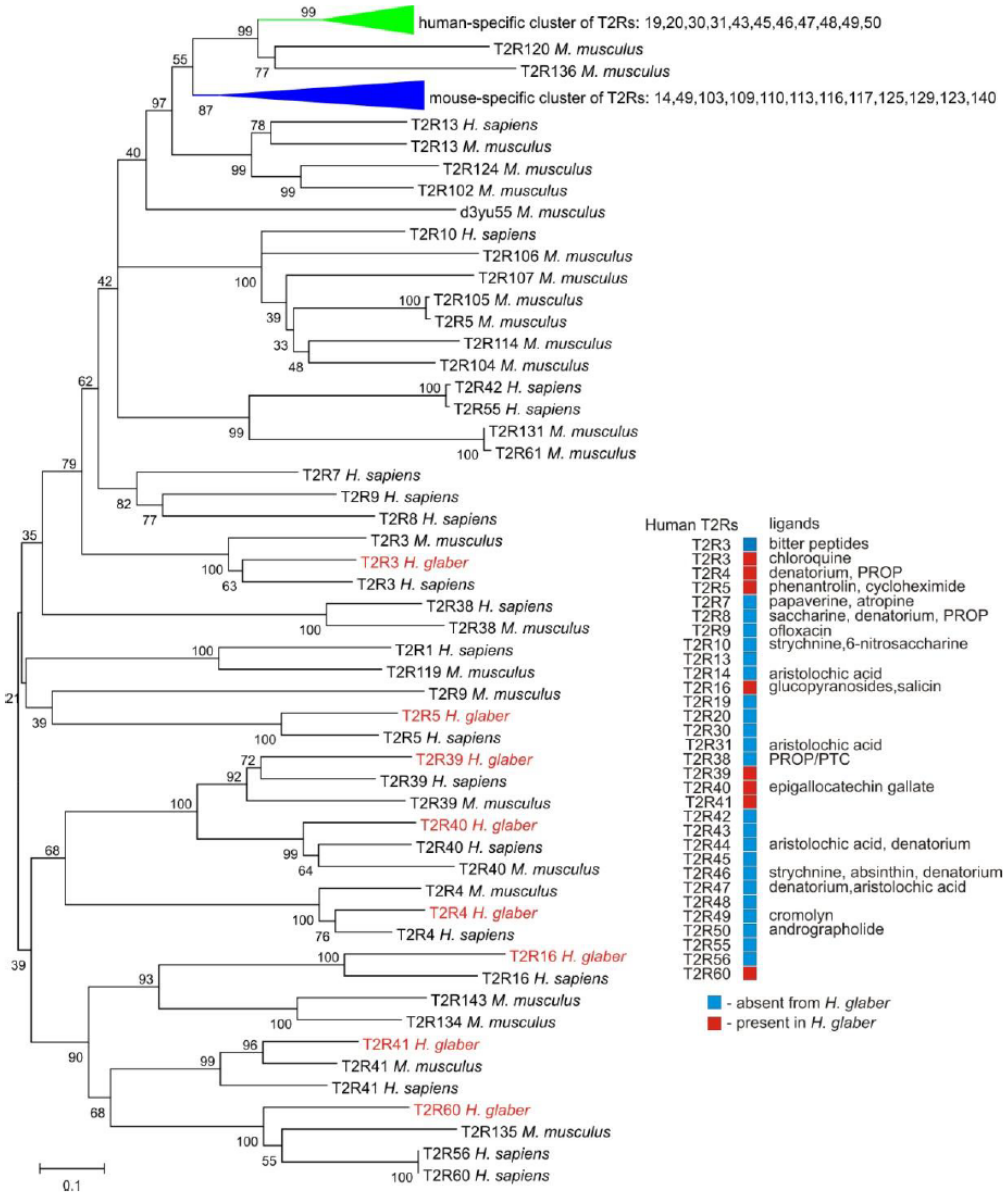
GO_ID	GO_Term	Adjusted p-value
GO:0004984	olfactory receptor activity	< 0.001
GO:0007601	visual perception	P=0.015
GO:0007283	spermatogenesis	P=0.044

POOR VISION, SMALL EYES: MULTIPLE MUTATIONS



Of the four vertebrate opsin genes (RHO, OPN1LW, OPN1MWand OPN1SW), two (OPN1LW and OPN1MW) were missing. However, the NMR has intact RHO (rhodopsin) and OPN4 (melanopsin), **supporting the presence of rod-dominated retinae and the capacity to distinguish light/dark cues**. Of about 200 genes associated with visual perception (GO:0007601) in humans and mice, almost 10% were inactivated or missing in the NMR.

BITTER TASTE RECEPTORS:



Supplementary Fig. 26.
The Neighbor-Joining phylogenetic tree demonstrating the relationships between eight NMR T2R proteins (in red) and known T2R proteins of human and mouse.

The evolutionary distances were computed using the Poisson correction method and are in the units of the number of amino acid substitutions per site. All positions containing gaps and missing data were eliminated from the dataset.

Unique AA or DNA changes

THERMOREGULATION: 4 amino acid changes in UCP1 gene

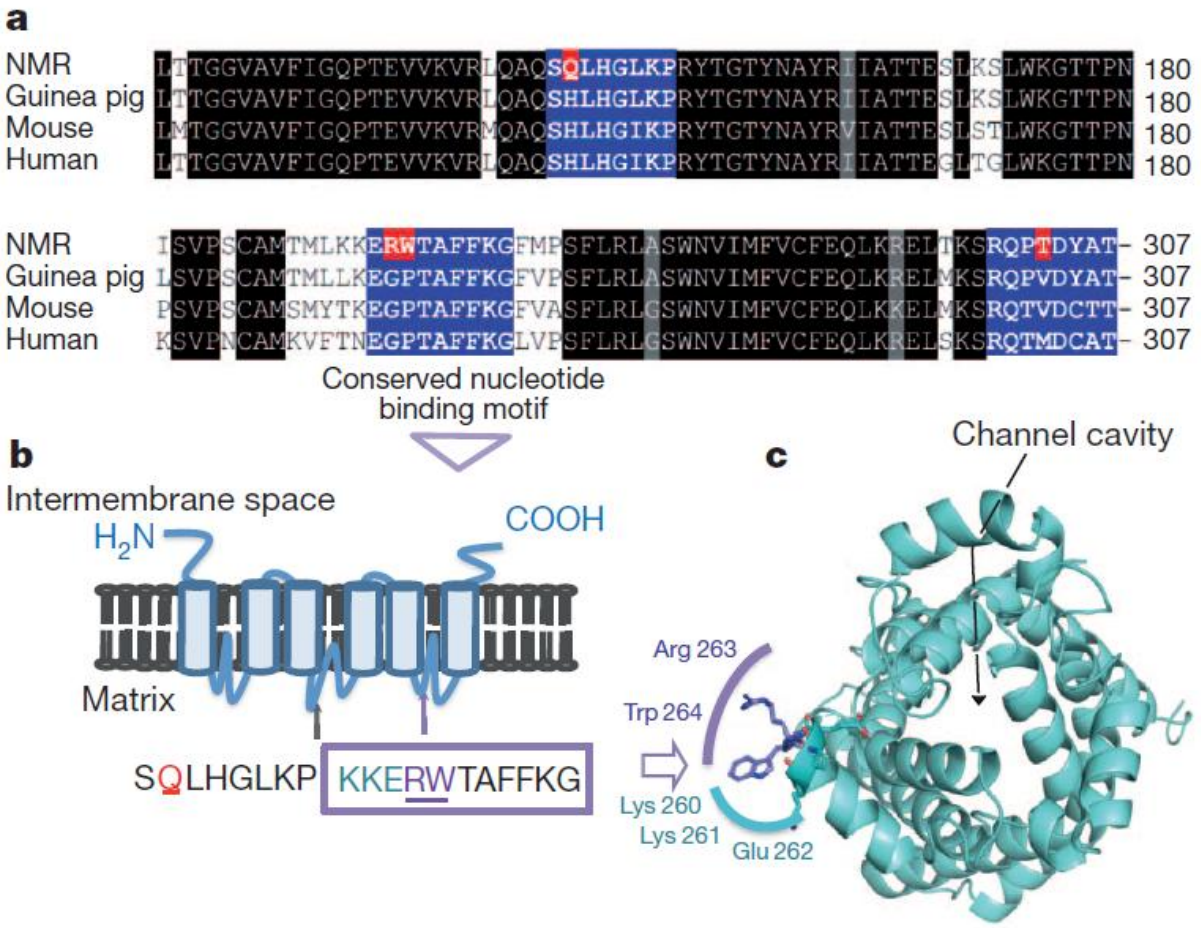
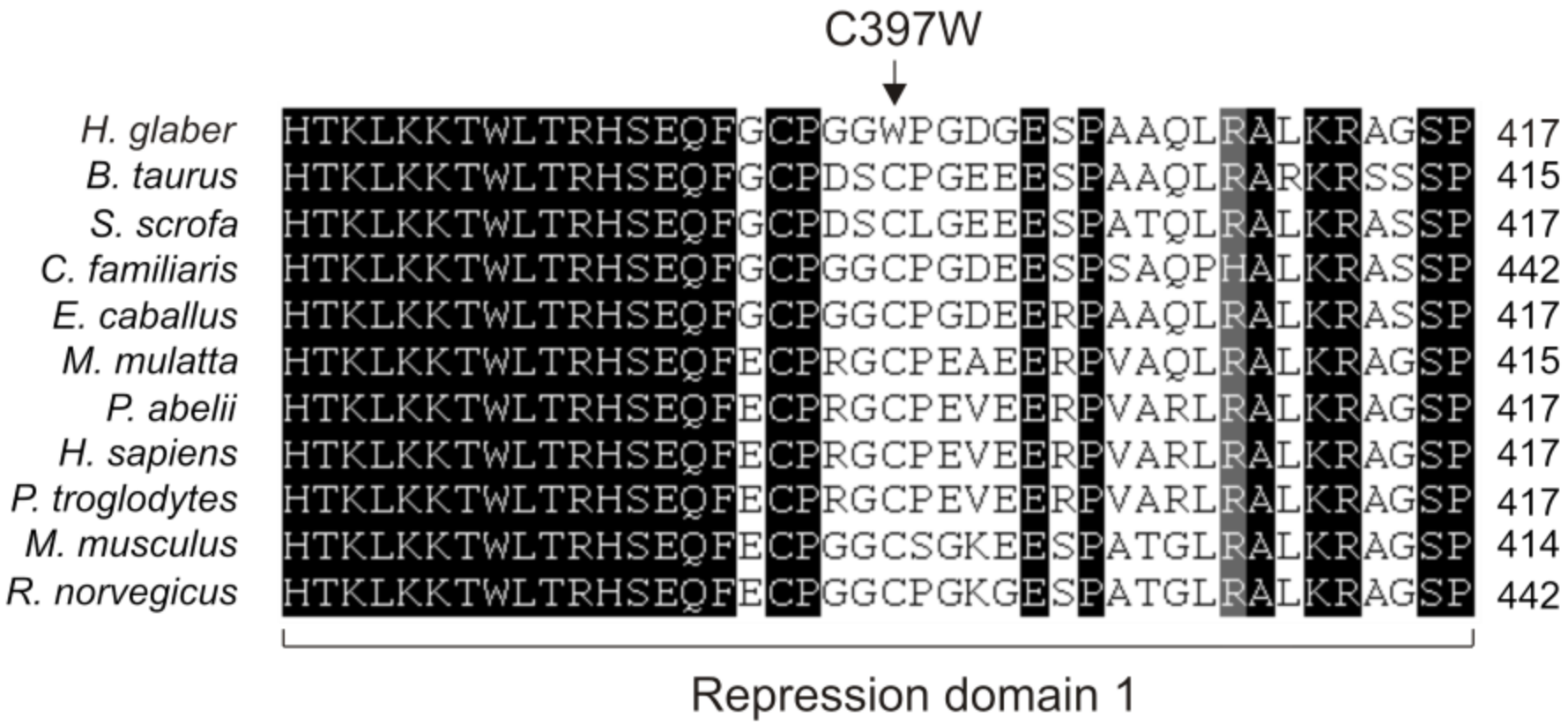


Figure 3. Unique changes in UCP1 sequences and their roles in thermoregulation.
a) Alignment of mammalian UCP1 sequences. Amino acids unique to the NMR are highlighted in red, and conserved motifs in blue. b) Topology of UCP1. Regions affected in the NMR are highlighted. c) Structural model of UCP1. Location of the channel and the nucleotide binding loop with altered sequences in the NMR are shown.

Unique AA or DNA changes

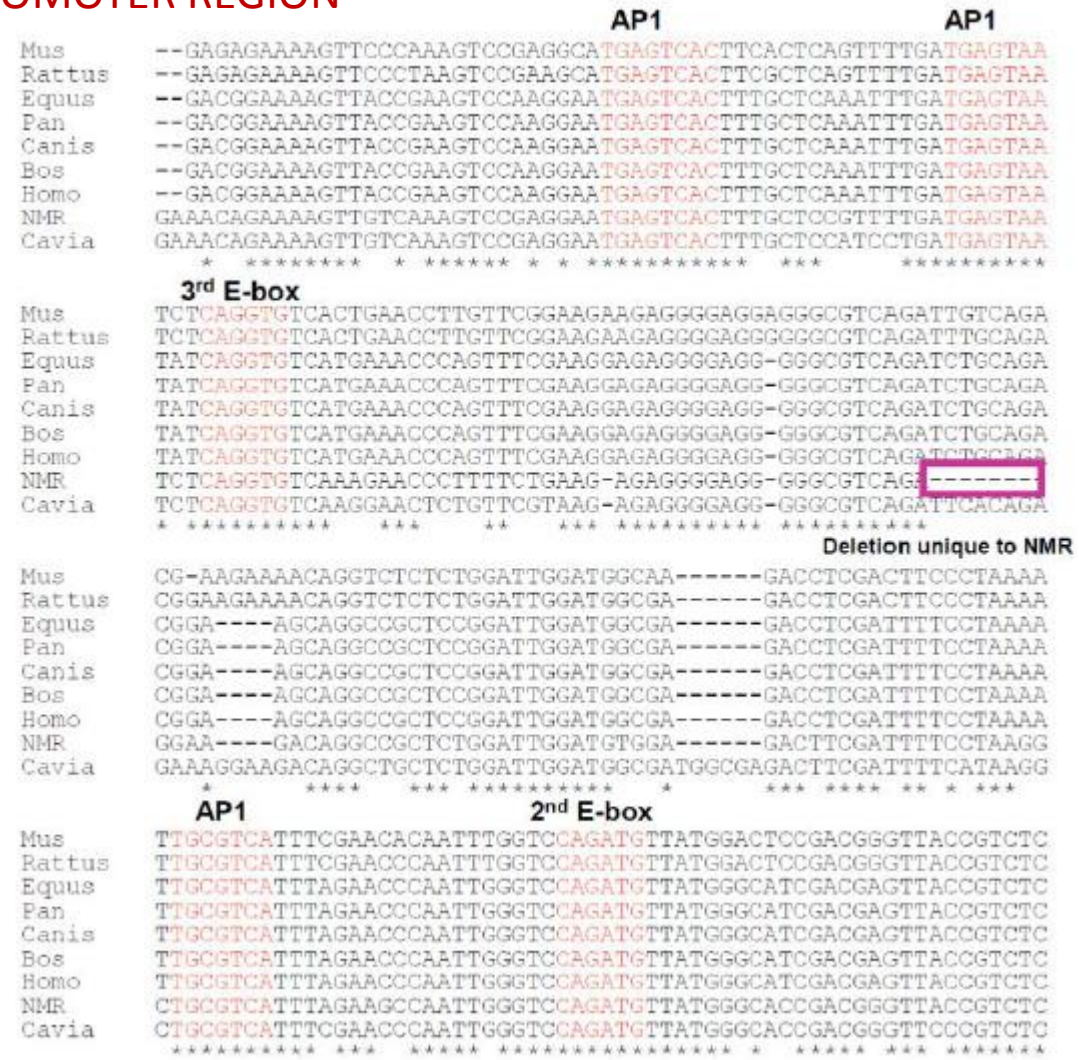
HAIRLESS: One Amino Acid change



Hairless rats have mutations in the same region C397Y and C422Y

Unique AA or DNA changes

PAIN SENSITIVITY: DELETION IN PROMOTER REGION



Supplementary Fig 22. NMR-specific deletion within the *TAC1* promoter. Transcription start sites of human and mouse genes are indicated with arrows. NMR-specific deletion within the *TAC1* promoter is indicated with a box. AP1 and E-box are transcription factor binding sites known to regulate *TAC1* expression.

Positive selection genes

AGING?: TELOMERASES

COL4A2	Collagen alpha-2(IV) chain	HIVEP2	Transcription factor HIVEP2
CCDC162	Coiled-coil domain-containing protein 162	TBR1	T-box brain protein 1
PCDHA3	Protocadherin alpha-3	BTF3	Transcription factor BTF3
RHOBTB2	**Rho-related BTB domain-containing protein 2**	**NCKAP5L**	**Nck-associated protein 5-like**
ROBO4	**Roundabout homolog 4**	KIAA0319	Dyslexia-associated protein
PEAR1	Platelet endothelial aggregation receptor 1	**PAK7**	**Serine/threonine-protein kinase PAK 7**
C1orf173	Uncharacterized protein C1orf173	ZNRD1-AS	Putative uncharacterized protein
TMPO	Lamina-associated polypeptide 2	**DNAJC1**	**DnaJ homolog subfamily C member 1**
ZNF167	Zinc finger protein 167	**TEP1**	**Telomerase protein component 1**
FLG2	Filaggrin-2	SLC19A3	Thiamine transporter 2
ABCA9	ATP-binding cassette subfamily A member 9	ABCC10	Multidrug resistance-associated protein 7
CARD6	**Caspase recruitment domain-containing protein 6**	OR56A3	Olfactory receptor 56A3
MEGF6	Multiple epidermal growth factor-like domains protein 6	**RPRD1A**	**Regulation of nuclear pre-mRNA domain-containing protein 1A**
CCDC15	Coiled-coil domain-containing protein 15	COL24A1	Collagen alpha-1(XXIV) chain
FGFR2	Fibroblast growth factor receptor 2	KCNQ1	Potassium voltage-gated channel subfamily KQT member 1
C12orf43	Uncharacterized protein C12orf43	COL3A1	Collagen alpha-1(III) chain
PCDHAC2	Protocadherin alpha-C2	MYL6	Myosin light polypeptide 6
C12orf43	Uncharacterized protein C1orf168	DMRTA2	Doublesex- and mab-3-related transcription factor A2
DPEP1	**Dipeptidase 1**	**E2F4**	**Transcription factor E2F4**
TAAR2	Trace amine-associated receptor 2	**OLFM4**	**Olfactomedin-4**
PCDHGB1	Protocadherin gamma-B1	CCDC27	Coiled-coil domain-containing protein 27
C2orf71	Uncharacterized protein C2orf71	GPR112	Probable G-protein coupled receptor 112
SLC9A11	**Sodium/hydrogen exchanger 11**		

Supplementary Table 18. Positively selected genes. 141 genes were identified by PAML's branch-site test of positive selection. Among the first 45 genes (with FDR<0.01), the genes shown in bold were checked manually. Some of the genes in this table, especially those not shown in bold, may be false-positives.

AGING and CANCER RESISTANCE: No good explanation

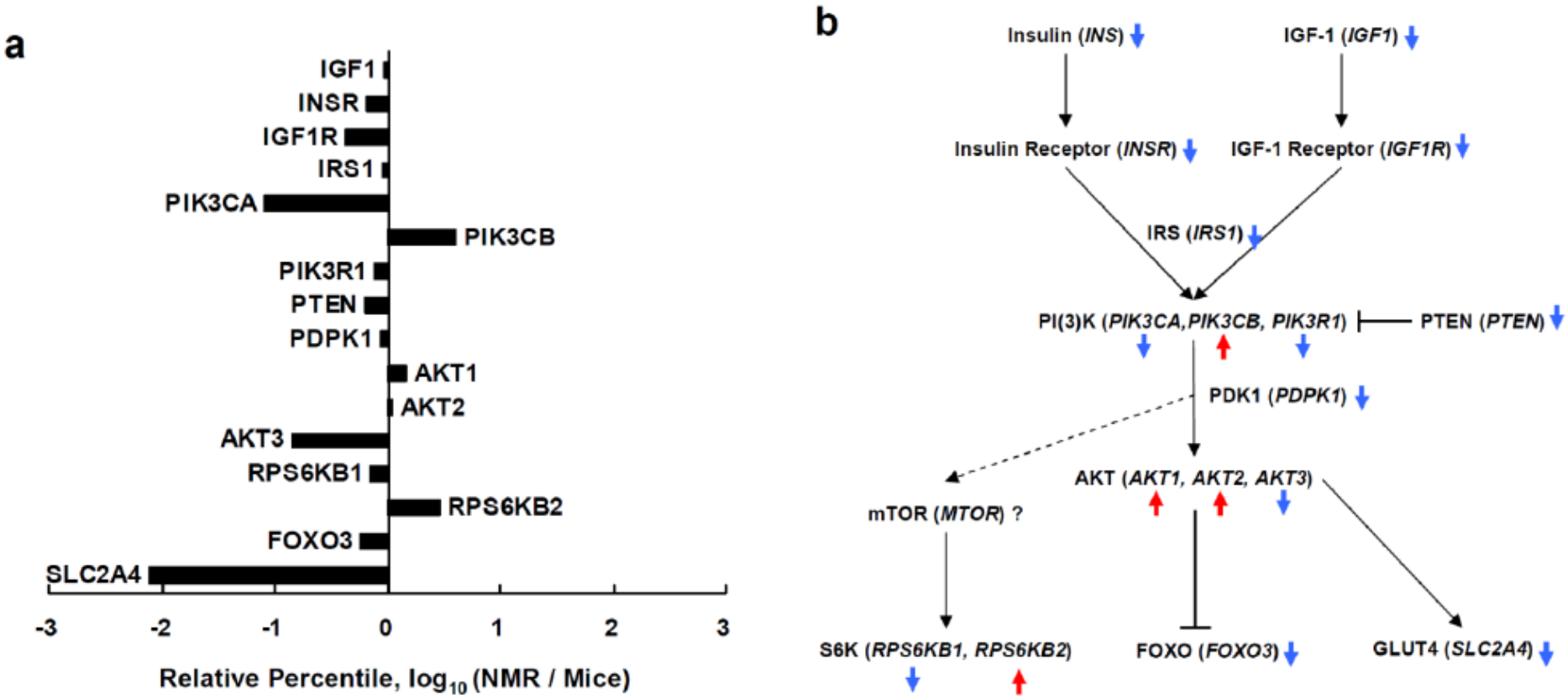
Age related gene expression in human was not observed for the same genes in NMT
Difficult to interpret (cause, consequence or just random noise of experiment)?

For example, genes related to degradation of macromolecules, such as GSTA1, DERL1 and GNS, were not upregulated with age inNMRs. We also found that genes encoding mitochondrial proteins (NDUFB11, ATP5G3 and UQCRQ) were not downregulated, consistent with stable maintenance of mitochondrial function during ageing. It is also of interest that TERT (telomerase reverse transcriptase) showed stable expression regardless of age.

Likewise, our transcriptome analysis of the NMR revealed decreased expression of genes involved in insulin/IGF-1 signalling in the liver compared to mice.

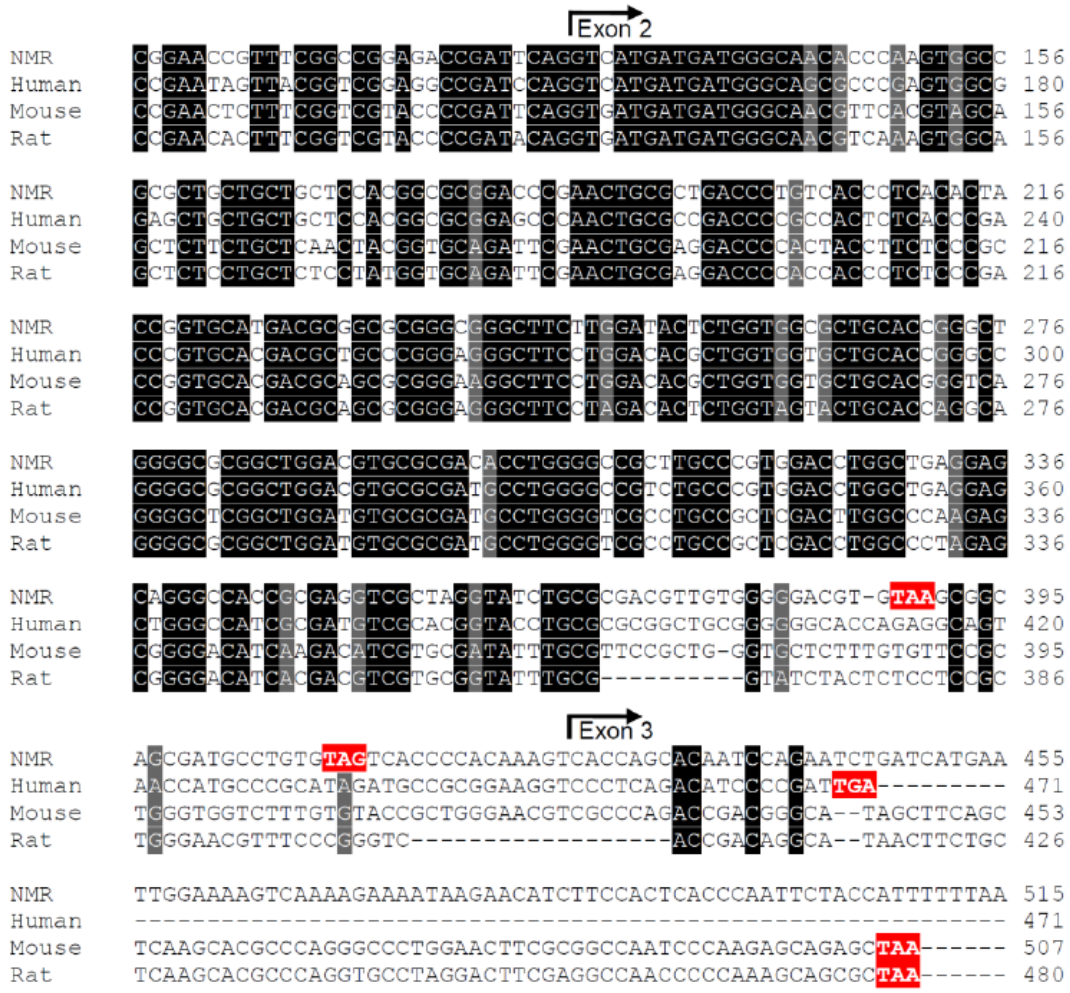
AGING and CANCER RESISTANCE: No good explanation

Likewise, our transcriptome analysis of the NMR revealed decreased expression of genes involved in insulin/IGF-1 signalling in the liver compared to mice.



AGING and CANCER RESISTANCE: No good explanation

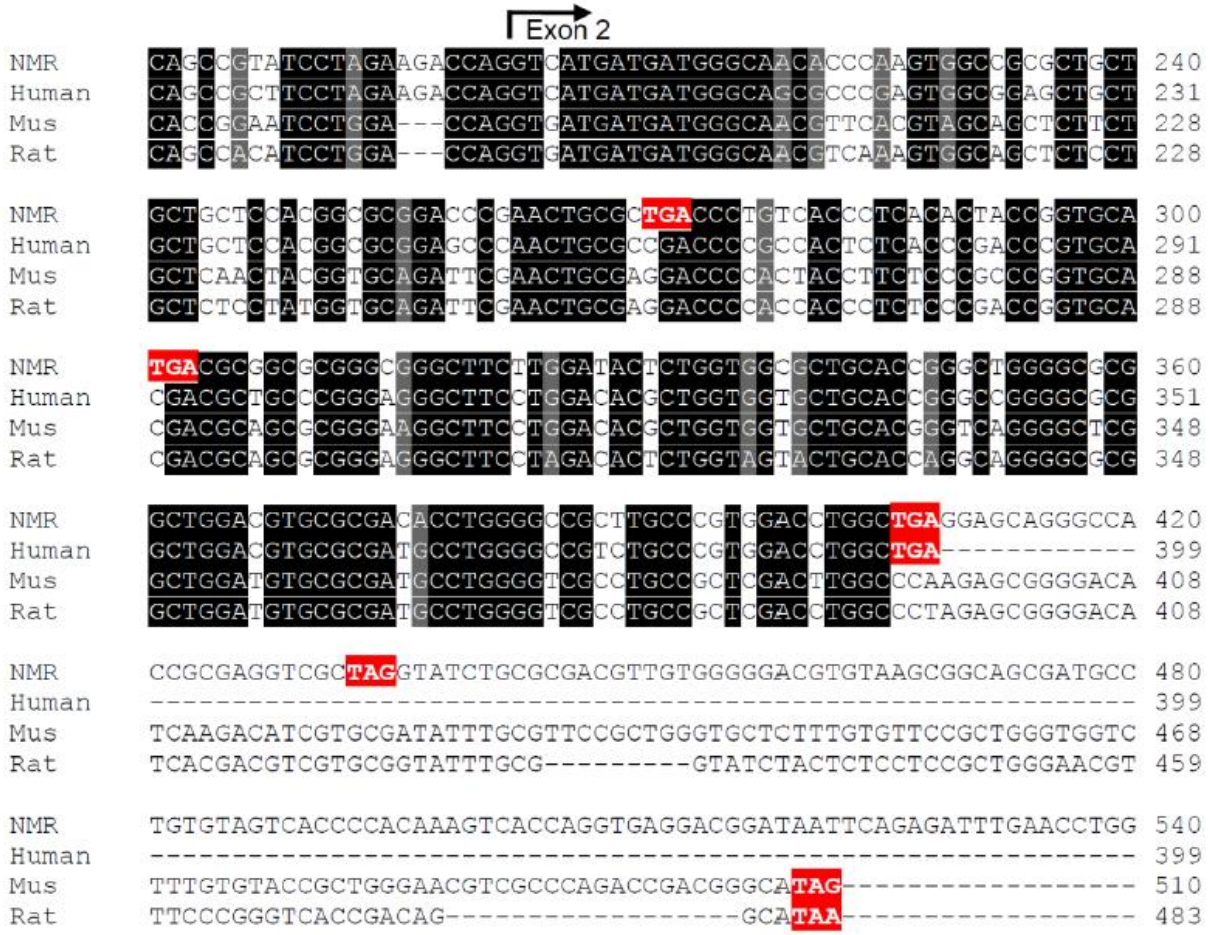
To explain the extraordinary resistance of the NMR to cancer, a two-tier protective mechanism involving contact inhibition mediated by $p16^{Ink4a}$ and $p27^{Kip1}$ was proposed. The involvement of $p16^{Ink4a}$ is unusual, since humans and mice show only contact inhibition mediated by $p27^{Kip1}$.



Alignment of mammalian Ink4a ($p16^{Ink4a}$) coding regions

AGING and CANCER RESISTANCE: No good explanation

To explain the extraordinary resistance of the NMR to cancer, a two-tier protective mechanism involving contact inhibition mediated by $p16^{Ink4a}$ and $p27^{Kip1}$ was proposed. The involvement of $p16^{Ink4a}$ is unusual, since humans and mice show only contact inhibition mediated by $p27^{Kip1}$.



Alignment of mammalian Arf ($p16^{Arf}$) coding regions