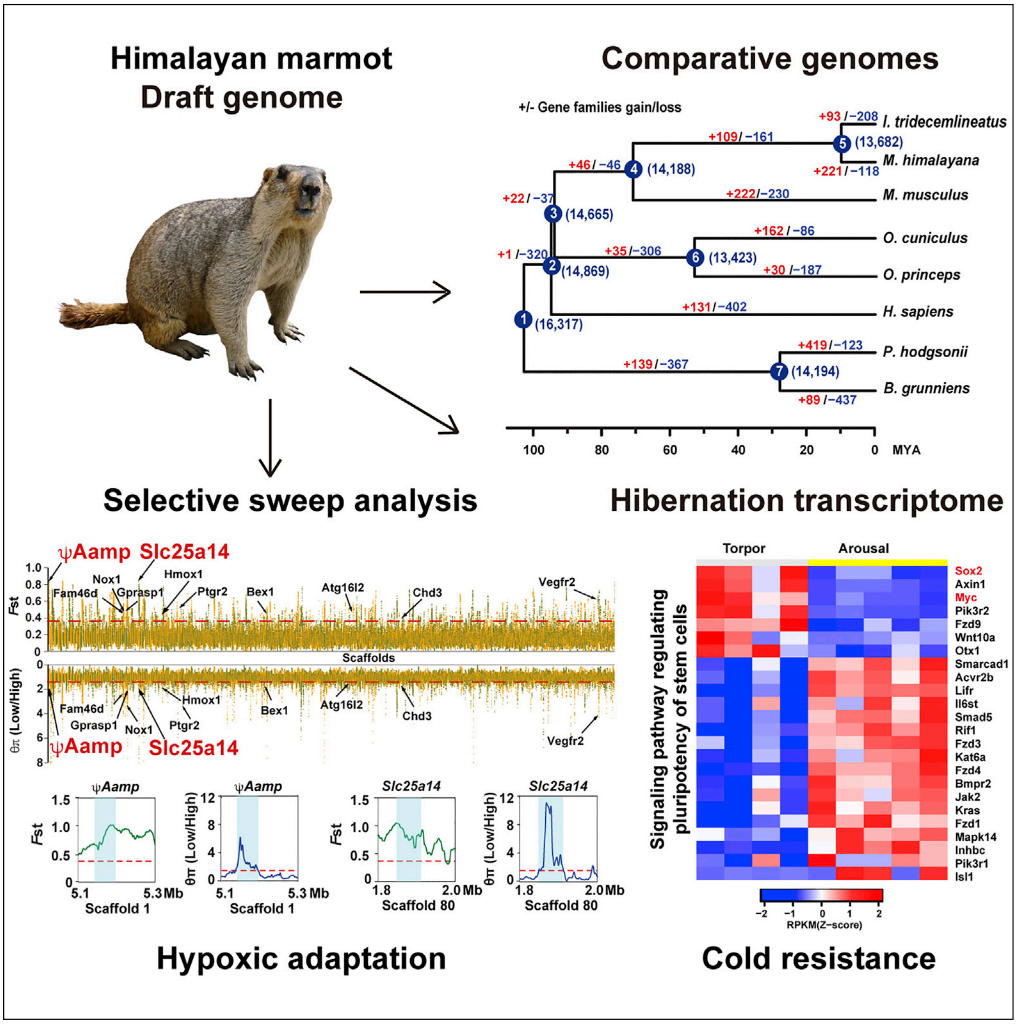
iScience

CellPress

Article

Hypoxic and Cold Adaptation Insights from the Himalayan Marmot Genome



Liang Bai, Baoning Liu, Changmian Ji, ..., Hongkun Zheng, Jianglin Fan, Enqi Liu

zhenghk@biomarker.com.cn (H.Z.)
jianglin@yamanashi.ac.jp (J.F.)
liuenqi@mail.xjtu.edu.cn (E.L.)

HIGHLIGHTS

This is the first report on a draft genome of Himalayan marmot

Selective sweep genes Slc25a14 and ψAamp were potentially involved in hypoxic adaptation

Stem cell pluripotency pathway may be implicated in cold resistance during hibernation

Bai et al., iScience *11*, 519–530
January 25, 2018 © 2018 The Author(s).
https://doi.org/10.1016/j.isci.2018.11.034

iScience

CellPress

Article

Hypoxic and Cold Adaptation Insights from the Himalayan Marmot Genome

Liang Bai,[1,2,11] Baoning Liu,[1,2,11] Changmian Ji,[3,11] Sihai Zhao,[1,2,11] Siyu Liu,[4] Rong Wang,[1,2] Weirong Wang,[1,2] Pu Yao,[3] Xuming Li,[3] Xiaojun Fu,[3] Haiyan Yu,[3] Min Liu,[3] Fengming Han,[3] Ning Guan,[3] Hui Liu,[3] Dongyuan Liu,[3] Yuanqing Tao,[5] Zhongdong Wang,[5] Shunsheng Yan,[6] Greg Florant,[7] Michael T. Butcher,[8] Jifeng Zhang,[9] Hongkun Zheng,[3,*] Jianglin Fan,[10,*] and Enqi Liu[1,2,12,*]

SUMMARY

The Himalayan marmot (*Marmota himalayana*) is a hibernating mammal that inhabits the high-elevation regions of the Himalayan mountains. Here we present a draft genome of the Himalayan marmot, with a total assembly length of 2.47 Gb. Phylogenetic analyses showed that the Himalayan marmot diverged from the Mongolian marmot approximately 1.98 million years ago. Transcriptional changes during hibernation included genes responsible for fatty acid metabolism in liver and genes involved in complement and coagulation cascades and stem cell pluripotency pathways in brain. Two selective sweep genes, Slc25a14 and ψAamp, showed apparent genotyping differences between low- and high-altitude populations. As a processed pseudogene, ψAamp may be biologically active to influence the stability of Aamp through competitive microRNA binding. These findings shed light on the molecular and genetic basis underlying adaptation to extreme environments in the Himalayan marmot.

INTRODUCTION

The Himalayan marmot (*Marmota himalayana*), a large squirrel of the genus *Marmota*, is widely distributed at elevations of 1,900–5,000 m around the Himalayan regions of India, Nepal, and Pakistan, and the Qinghai-Tibetan Plateau of China (Shrestha, 2016; Nikol'skii and Ulak, 2006). The Qinghai-Tibetan Plateau is known for its extreme environment with low atmospheric oxygen pressure, cold climate, and limited resources (Wu, 2001). The Himalayan marmot possesses several distinctive biological features, such as hibernation, deep burrow excavation, thick fur, and increased size, which may be associated with its evolutionary responses to the selective pressures of its harsh environment (Cardini and O'Higgins, 2005; Matthews, 1971).

Numerous wild animals, such as the yak, Tibetan antelope, and Tibetan Mastiff inhabit the Qinghai-Tibetan Plateau. Recently, the mechanisms of their adaptation to high altitude have been of great interest. Although different species have experienced similar selection pressures, divergent adaptive pathways and related genes may be involved (Qiu et al., 2012; Ge et al., 2013; Gou et al., 2014). Distinct from other plateau mammals, Himalayan marmots hibernate in family groups during winter time. Hibernation burrows are especially deep, in some cases over 10 m deep (Smith et al., 2010). Hibernation in mammals is a seasonal state of metabolic suppression and dormancy characterized by a decrease in body temperature, metabolism, heart rate, and oxygen consumption (Geiser, 2013). Himalayan marmots are confronted with severe hypoxic and cold stress during winter.

Here, we report the deep sequencing and *de novo* assembly of a male Himalayan marmot and resequencing of 20 Himalayan marmots from high- and low-altitude and 4 other marmot species, as well as RNA sequencing of Himalayan marmots from the torpor/arousal stage. This study provides clues regarding the genetic mechanisms underlying high-altitude adaptation and hibernation, and will be a valuable resource for researchers studying *Marmota* evolution, highland disease, cold adaptation, and basic Himalayan marmot biology.

RESULTS

Genome Assembly and Annotation

Using a whole-genome shotgun strategy with the Illumina HiSeq 2500 platform, we sequenced the genome of a male Himalayan marmot from Xining, Qinghai province, China. The *de novo* assembly of a 508.41-Gb

[1] Laboratory Animal Center, Xi'an Jiaotong University Health Science Center, No.76, Yanta West Road, Xi'an, Shaanxi 710061, China

[2] Research Institute of Atherosclerotic Disease, Xi'an Jiaotong University Cardiovascular Research Center, Xi'an, Shaanxi 710061, China

[3] Biomarker Technologies Corporation, Beijing 101200, China

[4] Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA

[5] Qinghai Institute for Endemic Disease Prevention and Control, Xining, Qinghai 811602, China

[6] Centers for Disease Control and Prevention, Urumqi, Xinjiang 830054, China

[7] Department of Biology, Colorado State University, Ft. Collins, CO 80523, USA

[8] Department of Biological Sciences, Youngstown State University, Youngstown, OH 44555, USA

[9] Center for Advanced Models for Translational Sciences and Therapeutics, University of Michigan Medical Center, Ann Arbor, MI 48109, USA

[10] Department of Molecular Pathology, Faculty of Medicine, Interdisciplinary Graduate School of Medicine, University of Yamanashi, 1110, Shimokato, Chuo, Yamanashi 409-3898, Japan

[11] These authors contributed equally

[12] Lead Contact

Continued

high-quality sequence from 17 paired-end and mate-pair libraries provided 206-fold coverage with a total assembly length of 2.47 Gb, which approximates the genome size estimated by 21 K-mer distribution (Table S1; Figure S1). The contig N50 and scaffold N50 were 80 Kb and 1.5 Mb, respectively (Table S2). The high level of completeness and accuracy of the assembly were validated (Tables S3–S5; Figures S2 and S3). Repeat content comprised approximately 46.52% of the Himalayan marmot genome (Table S6), which was similar to that of the human genome at 46.1% (Li et al., 2010). The Himalayan marmot shares a similar profile of GC content and CpG frequency as the ground squirrel (Figure S4). A total of 4,905 non-coding RNAs and 1,479 pseudogenes were identified in the Himalayan marmot genome (Tables S7 and S8). Using a combinational method based on homology, *de novo*, and transcriptome data, 21,609 protein-coding genes were predicted, and 99.4% of them were functionally annotated (Tables S9–S11). Gene model features were similar to those of model animals, suggesting the high accuracy of the Himalayan marmot gene set (Table S12; Figure S5).

Genome Evolution

To ascertain the phylogenetic position of Himalayan marmot, we compared the genome coding sequence of the Himalayan marmot with those of 14 mammals spanning the orders Primates, Rodentia, Lagomorpha, Perissodactyla, and Artiodactyla. The phylogenetic tree placed the Himalayan marmot and ground squirrel in the same clade (Figure 1A). This result is consistent with the cladistics data that suggest that marmots evolved from ground squirrels (Hoffmann and Nadler, 1968; Thomas and Martin, 1993). The divergence time between the Himalayan marmot and ground squirrel was estimated to be ~9.8 million years ago (MYA) (Figure 1A). By comparing the protein sequences of Himalayan marmot to those of humans, mice, and rabbits, we identified 10,358 homologous gene families shared by the four mammals and 235 gene families that were specific to Himalayan marmot (Figure S6). These families were significantly overrepresented in the categories of regulation of feeding behavior and fatty acid metabolic process (corrected $p < 0.05$) (Table S13).

Variance of gene number in each family has been proposed as a major mechanism underlying the adaptive divergence of closely related species. Compared with ground squirrel, we inferred 221 and 118 gene families that were substantially expanded and contracted in the Himalayan marmot, respectively (Figure 1B). The expanded gene families were mainly enriched in the functional categories of feeding behavior (GO:2000253, positive regulation of feeding behavior; GO:0002021, sensory perception of smell; etc.; corrected $p < 0.05$), hypoxic adaptation (GO:0003300, cardiac muscle hypertrophy; GO:0007596, blood coagulation; etc.; corrected $p < 0.05$), and energy metabolism (GO:2000507, positive regulation of energy homeostasis; etc.; corrected $p < 0.05$) (Table S14). Based on the hypothesis that the rapidly evolving genes have been under positive selection, we identified 78 positively selected genes (PSGs) using the branch-site likelihood ratio test (Table S15). These PSGs were regarded to be involved in the functional categories of regulation of systemic arterial blood pressure (GO:0043281), G2 DNA damage checkpoint (GO:0031572), triglyceride metabolic processes (GO:0006641), cardiac muscle contraction (GO:0060048), etc. (Table S16). The specific and expanded gene families and PSGs in Himalayan marmot infer that adaptive evolution to the harsh environment occurred at the genomic level.

The Divergence and Demographic History of *Marmota*

To elucidate the evolutionary scenarios of *Marmota* species from a genome-wide perspective, we sampled four *Marmota* species, including Mongolian marmot (*Marmota sibirica*), gray marmot (*Marmota baibacina*), long-tailed marmot (*Marmota caudate*), and yellow-bellied marmot (*Marmota flaviventris*), and performed whole genome resequencing with approximately 10-fold coverage for each individual (Figure 2A; Table S17). Phylogenetic analysis showed that the Himalayan marmot and Mongolian marmot are sister species and that their divergence time is ~1.98 MYA (Figure 2B). As sister species, the heterozygosity rates of Himalayan marmot and Mongolian marmot were comparable (Figure 2C). However, the historical trends of effective population size (N_e) for Himalayan marmot and Mongolian marmot showed a quite different pattern (Figure S7), suggesting that their adaptation to their specific habitats has occurred independently and that these adaptations relate to a number of facets of their biology, including seasonality, food habits, and social behavior. The heterozygosity in yellow-bellied marmot was higher than in other *Marmota* species (Figure 2C), coinciding with its wide distribution and large population size based on the pairwise sequentially Markovian coalescent (PSMC) model (Figure S7) (Cardini, 2003). Unexpectedly, the speciation of the Alpine marmot (*Marmota marmota*) (~5.45 MYA) was earlier than that of the yellow-bellied marmot based on whole-genome sequence and Y chromosome

*Correspondence:
zhenghk@biomarker.com.cn
(H.Z.),
jianglin@yamanashi.ac.jp
(J.F.),
liuenqi@mail.xjtu.edu.cn
(E.L.)

https://doi.org/10.1016/j.isci.2018.11.034

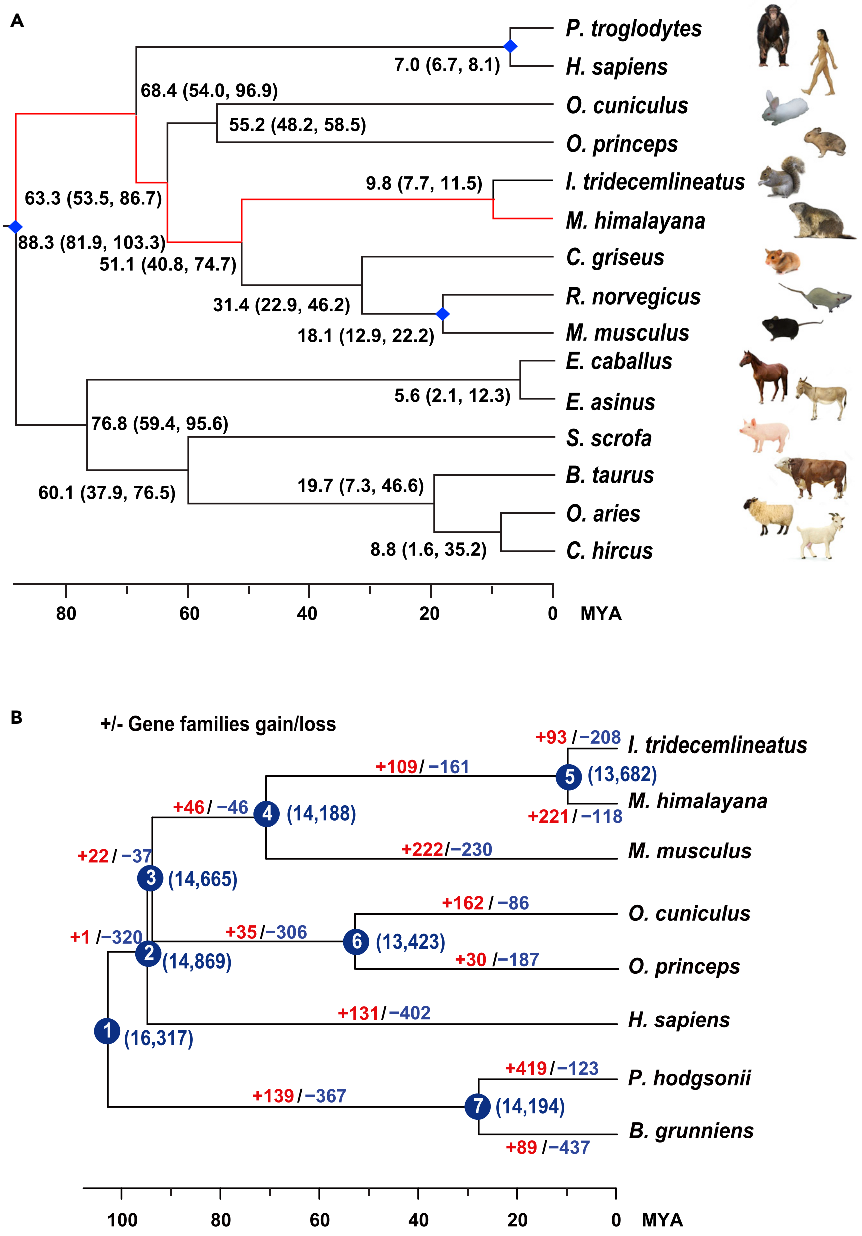


Figure 1. Genome Evolution of the Himalayan Marmot

(A) Phylogenetic tree of 15 mammals constructed by the maximum likelihood method. The branch of the Himalayan marmot is highlighted in red. The divergence time was estimated using the nodes with calibration times derived from the Time Tree database, which were marked by a blue rhombus. All estimated divergence times are shown with 95% confidence intervals in brackets.

(B) Gene expansion and contraction in the Himalayan marmot genome. The number of expanded (red) and contracted (blue) gene families are shown along branches and nodes. The blue digits in the parentheses represent the estimated numbers of gene families in the common ancestral species. MYA, million years ago. *P. troglodytes, Pan troglodytes; H. sapiens, Homo sapiens; O. cuniculus, Oryctolagus cuniculus; O. princeps, Ochotona princeps; I. tridecemlineatus, Ictidomys tridecemlineatus; M. himalayana, Marmota himalayana; C. griseus, Cricetulus griseus; R. norvegicus, Rattus norvegicus; M. musculus, Mus musculus; E. caballus, Equus caballus; E. asinus, Equus asinus; S. scrofa, Sus scrofa; B. taurus, Bos taurus; O. aries, Ovis aries; C. hircus, Capra hircus; P. hodgsonii, Pantholops hodgsonii; B. grunniens, Bos grunniens.*

See also Figures S1–S6 and Tables S1–S7, S8, S9–S11, S12, S13, S14, S15, and S16.

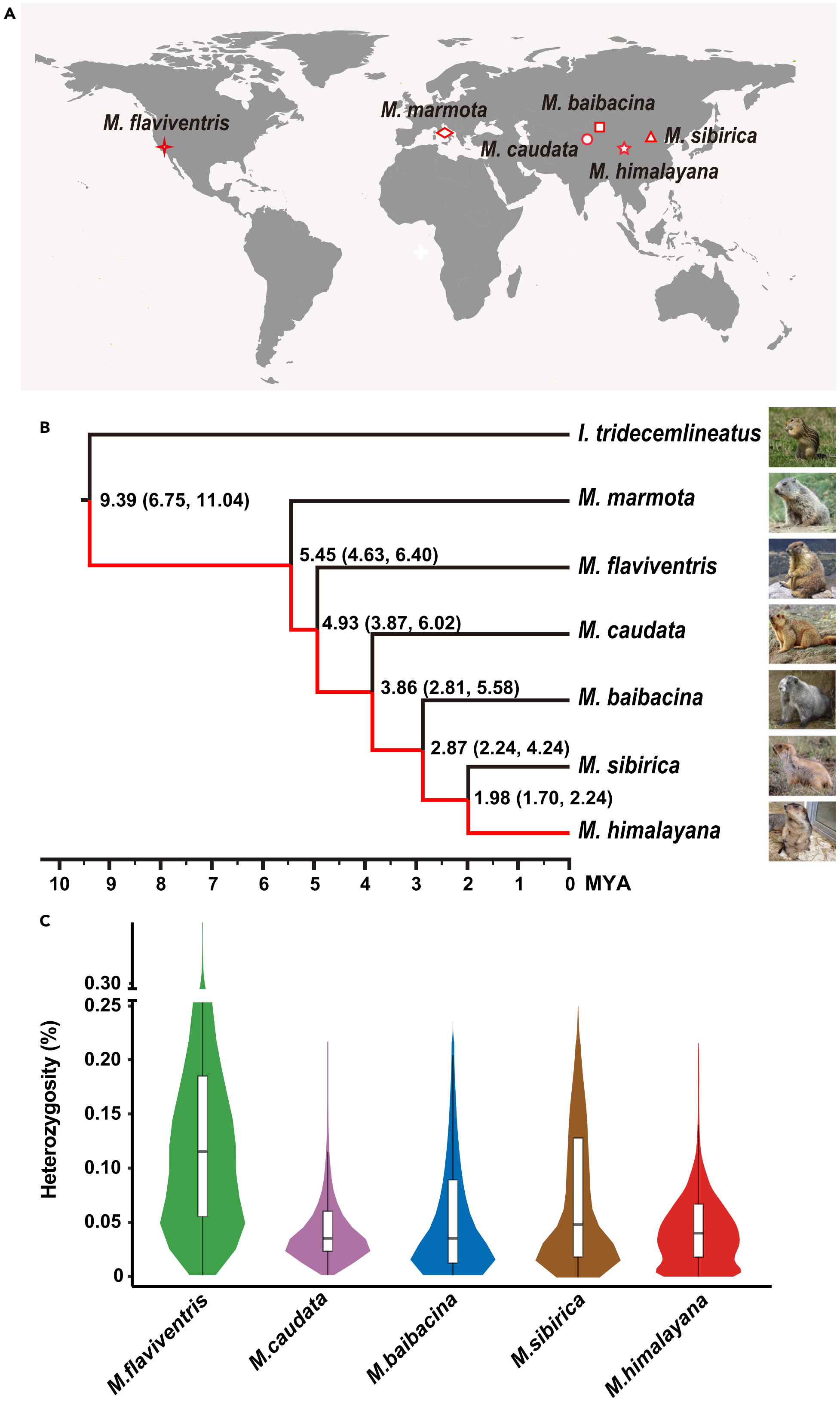

A
M. flaviventris
M. marmota
M. baibacina
M. sibirica
M. caudata
M. himalayana
B
I. tridecemlineatus
M. marmota
M. flaviventris
M. caudata
M. baibacina
M. sibirica
M. himalayana
9.39 (6.75, 11.04)
5.45 (4.63, 6.40)
4.93 (3.87, 6.02)
3.86 (2.81, 5.58)
2.87 (2.24, 4.24)
1.98 (1.70, 2.24)
10
9
8
7
6
5
4
3
2
1
0
MYA
C
Heterozygosity (%)
0.30
0.25
0.20
0.15
0.10
0.05
0
M.flaviventris
M.caudata
M.baibacina
M.sibirica
M.himalayana

Figure 2. Geographic Locations, Phylogenetic Relationships, and Heterozygosity of *Marmota* Species

(A) Sampling localities of six *Marmota* species. Quadrangle, *M. flaviventris*, yellow-bellied marmot; diamond, *M. marmota,* alpine marmot; square, *M. baibacina*, gray marmot; circle, *M. caudate*, long-tailed marmot; triangle, *M. sibirica*, Mongolian marmot; star, *M. himalayana,* Himalayan marmot.

(B) Phylogenetic relationships of *Marmota* species. The branch of the Himalayan marmot is highlighted in red. All estimated divergence times are shown with 95% confidence intervals in brackets. MYA, million years ago.

(C) Heterozygosity of *Marmota* species. Each ''violin'' contains a white box (25%–75% range) and a horizontal black line (median), with the width denoting a kernel density trace.

See also Figures S7–S9, and Table S17.

gene sequence (Figures 2B and S8), challenging the contention that *Marmota* may originate from North America (Steppan et al., 1999).

Next, we applied the PSMC model to examine the changes in N_e of the ancestral population of four ''Himalayan'' species, including Himalayan marmot, giant panda, yak, and snub-nosed monkey. The N_e of the Himalayan marmot sharply declined during the two largest Pleistocene glaciations: the Xixiabangma glaciation (1.17∼0.8 MYA) and Naynayxungla glaciation (0.78–0.50 MYA) (Zheng et al., 2002) (Figure S9). Population reductions at these times also occurred in three other Himalayan species (Qiu et al., 2015; Zhao et al., 2013; Zhou et al., 2014) (Figure S9). Subsequently, Himalayan marmots underwent a much longer bottleneck period compared with other mammals (Figure S9), which might be due to the relatively small body size of Himalayan marmots. However, the N_e of Himalayan marmots was not seriously affected by the last glacial maximum (LGM, ∼20,000 years ago), suggesting that Himalayan marmots had already adapted to the harsh environment (Figure S9).

Transcriptomic Analyses of Hibernation

In contrast to other plateau mammals, the Himalayan marmot hibernates during winter months. To elucidate the molecular mechanism underlying hibernation, we analyzed the RNA sequencing data derived from liver and brain to characterize the variations in gene expression during the torpor/arousal cycle. Differentially expressed genes (DEGs) were significantly enriched in the pathways of fatty acid metabolism, terpenoid backbone biosynthesis, and primary bile acid biosynthesis in liver ($p < 0.05$; Figure 3A). We clearly observed the upregulation of genes participating in fatty acid degradation (Hadha, Ehhadh, Acat1, Acdvl, etc.) and the downregulation of genes involved in fatty acid synthesis (Fasn, Acaca, Scd1, Elovl6, etc.) in the torpor state, suggesting a precise regulation of lipid-based metabolism (Figure 3B). In addition, we found an overall downregulation of genes involved in drug metabolism-cytochrome P450, biosynthesis of amino acids, and carbohydrate catabolism relative to arousal stage (Figure S10).

In the brain, DEGs were mainly enriched in the pathways of complement and coagulation cascades and signaling pathways regulating the pluripotency of stem cells, etc. ($p < 0.05$; Figure 3C). A previous report indicated that because the hibernator's brain is exposed to near-freezing temperatures and has decreased blood flow, there is increased risk of stasis-induced blood clots (Laursen et al., 2015). We observed significant downregulation of genes involved in the complement and coagulation cascades, metabolism of xenobiotics by cytochrome P450, and circadian rhythm (Figure S11). Moreover, a recent study demonstrated that neurons differentiated from hibernator's induced pluripotent stem cells retain intrinsic cold-resistant features (Ou et al., 2018). Two master transcription factors, Sox2 and Myc, in signaling pathways regulating the pluripotency of stem cells were remarkably upregulated during torpor (Figure 3D), which maintains the self-renewal capacity of stem cells and protects the brain from cold-induced injury (Takahashi and Yamanaka, 2006). Meanwhile, the activation of genes (Lifr, Bmpr2, Acvr2b, etc.) involved in regulating the pluripotency of stem cells during arousal (Figure 3D) may promote stem cell differentiation to repair injured cells. The temporal regulation of the pluripotency of stem cells may be a remarkable strategy for Himalayan marmots to survive extreme environmental stresses.

According to previous studies, Himalayan marmot shares common pathways with ground squirrel and black bear, which include lipid and glucose metabolism, detoxification, complement and coagulation cascades, and circadian rhythm (Seim et al., 2013; Fedorov et al., 2011; Williams et al., 2005). However, we found a small fraction of the predicted shared DEGs and paradoxical expression pattern of several genes between these two hibernators and the Himalayan marmot (Table S18; Figure S12). These results suggest

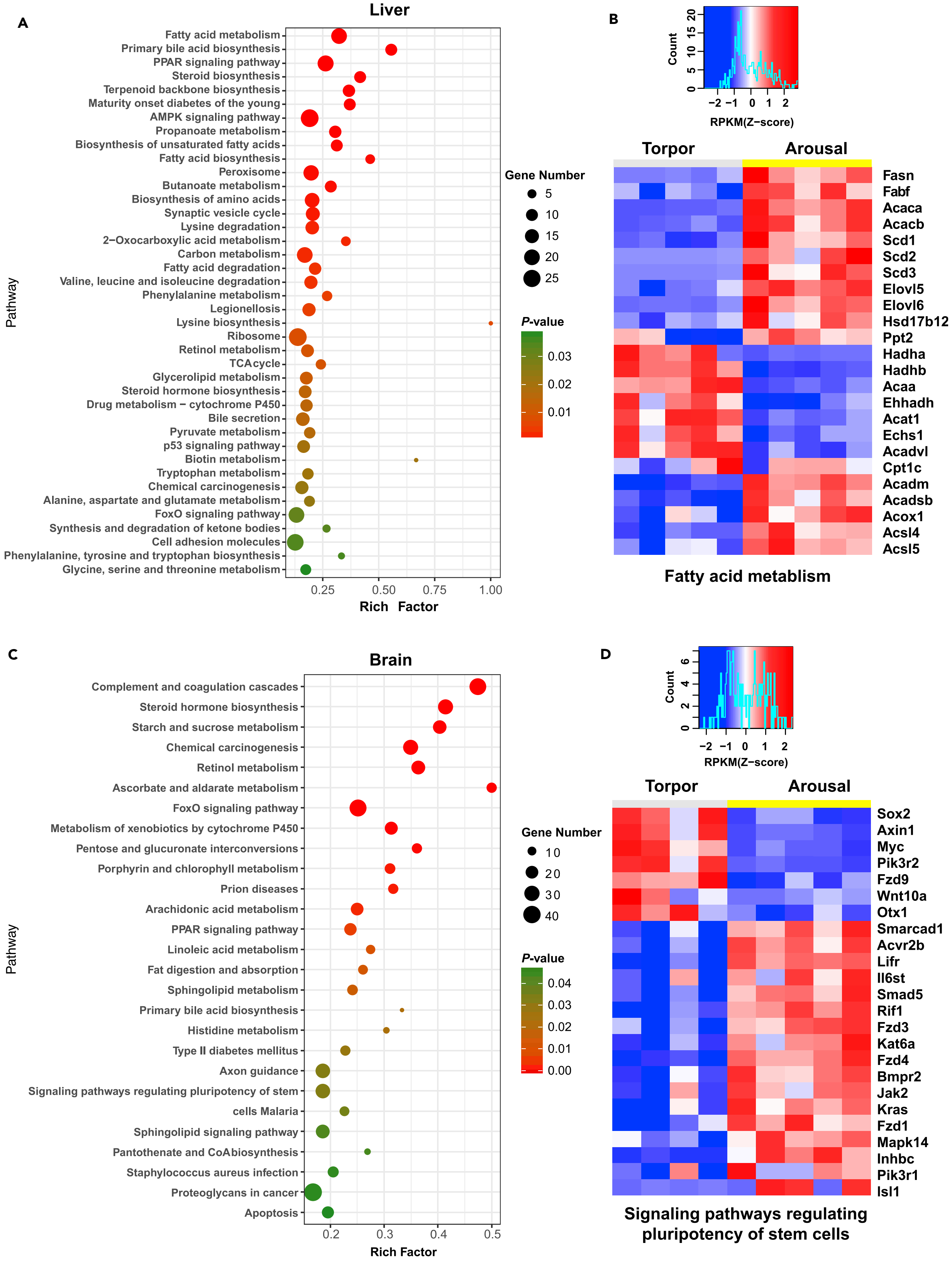

A
Liver
Pathway
Fatty acid metabolism
Primary bile acid biosynthesis
PPAR signaling pathway
Steroid biosynthesis
Terpenoid backbone biosynthesis
Maturity onset diabetes of the young
AMPK signaling pathway
Propanoate metabolism
Biosynthesis of unsaturated fatty acids
Fatty acid biosynthesis
Peroxisome
Butanoate metabolism
Biosynthesis of amino acids
Synaptic vesicle cycle
Lysine degradation
2−Oxocarboxylic acid metabolism
Carbon metabolism
Fatty acid degradation
Valine, leucine and isoleucine degradation
Phenylalanine metabolism
Legionellosis
Lysine biosynthesis
Ribosome
Retinol metabolism
TCAcycle
Glycerolipid metabolism
Steroid hormone biosynthesis
Drug metabolism − cytochrome P450
Bile secretion
Pyruvate metabolism
p53 signaling pathway
Biotin metabolism
Tryptophan metabolism
Chemical carcinogenesis
Alanine, aspartate and glutamate metabolism
FoxO signaling pathway
Synthesis and degradation of ketone bodies
Cell adhesion molecules
Phenylalanine, tyrosine and tryptophan biosynthesis
Glycine, serine and threonine metabolism
0.25
0.50
0.75
1.00
Rich Factor
Gene Number
5
10
15
20
25
P-value
0.03
0.02
0.01
B
Count
0 5 10 15 20
−2 −1 0 1 2
RPKM(Z−score)
Torpor
Arousal
Fasn
Fabf
Acaca
Acacb
Scd1
Scd2
Scd3
Elovl5
Elovl6
Hsd17b12
Ppt2
Hadha
Hadhb
Acaa
Ehhadh
Acat1
Echs1
Acadvl
Cpt1c
Acadm
Acadsb
Acox1
Acsl4
Acsl5
Fatty acid metablism
C
Brain
Pathway
Complement and coagulation cascades
Steroid hormone biosynthesis
Starch and sucrose metabolism
Chemical carcinogenesis
Retinol metabolism
Ascorbate and aldarate metabolism
FoxO signaling pathway
Metabolism of xenobiotics by cytochrome P450
Pentose and glucuronate interconversions
Porphyrin and chlorophyll metabolism
Prion diseases
Arachidonic acid metabolism
PPAR signaling pathway
Linoleic acid metabolism
Fat digestion and absorption
Sphingolipid metabolism
Primary bile acid biosynthesis
Histidine metabolism
Type II diabetes mellitus
Axon guidance
Signaling pathways regulating pluripotency of stem
cells Malaria
Sphingolipid signaling pathway
Pantothenate and CoAbiosynthesis
Staphylococcus aureus infection
Proteoglycans in cancer
Apoptosis
0.2
0.3
0.4
0.5
Rich Factor
Gene Number
10
20
30
40
P-value
0.04
0.03
0.02
0.01
0.00
D
Count
0 2 4 6
−2 −1 0 1 2
RPKM(Z−score)
Torpor
Arousal
Sox2
Axin1
Myc
Pik3r2
Fzd9
Wnt10a
Otx1
Smarcad1
Acvr2b
Lifr
Il6st
Smad5
Rif1
Fzd3
Kat6a
Fzd4
Bmpr2
Jak2
Kras
Fzd1
Mapk14
Inhbc
Pik3r1
Isl1
Signaling pathways regulating
pluripotency of stem cells

Figure 3. Characterization of Differentially Expressed Genes during Hibernation

(A–D) Distribution of hibernation-related differentially expressed genes (DEGs) in the liver (A) and brain (C) by pathway. Pathways were determined by searching the KEGG pathway database. The x axis indicates the richness factor (DEGs/background genes). Pathways with red and large dots are potentially important to the regulation of the torpor/arousal cycle. Heatmap constructed from liver (B) and brain (D) RNA sequencing data to characterize the gene variations in the pathways of fatty acid metabolism and stem cell pluripotency during torpor/arousal cycle.
See also Figures S10–S12, and Table S18.

that Himalayan marmots may utilize a diverse adaptive strategy to survive in highly seasonal or unpredictable environments.

Genetic Evolution for High-Altitude Adaptation

To better understand the mechanism of plateau adaptation for Himalayan marmot, we sampled 20 Himalayan marmots from extremely high-altitude (>4,500 m above sea level, n = 10) and relatively low- altitude (<1,900 m, n = 10) (Table S19). We first measured hematologic parameters and found that blood-related traits, especially red blood cell count, hemoglobin concentration, and mean corpuscular volume, were significantly increased in high-altitude Himalayan marmots compared with low-altitude Himalayan marmots (Table S20). Next, we performed whole-genome sequencing for each individual Himalayan marmot and searched the Himalayan marmot genome for regions with high population differentiation (F_{st}) and the ratio of pairwise diversity ($\theta_{\pi,\text{ low}}/\theta_{\pi,\text{ high}}$) (Table S19; Figures S13 and S14). We identified 24.84-Mb selective sweep regions distributed among different scaffolds of the Himalayan marmot genome. These regions contain 383 functional genes, which had significantly higher Ka/Ks value than nonselective genes ($p < 0.01$) (Tables S21 and S22; Figure S15). These genes are mainly classified into the categories of ''response to hypoxia,'' ''DNA repair,'' ''angiogenesis,'' ''heart function,'' ''fatty acid metabolism,'' ''cell cycle,'' ''heat generation,'' and ''calcium signaling pathway'' (Table S23). Ten functional genes involved in hypoxia-inducible factor-1 (HIF-1), vascular endothelial growth factor (VEGF), or other hypoxia-related signaling pathways, including Slc25a14, Nox-1, Hmox1, Vegfr2, Atg16l2, Bex1, Ptgr2, Gprasp1, Fam46d, and Chd3, showed clear genotyping differences between the high- and low-altitude groups (Figures 4A and S16; Table S23). A nonsynonymous substitution (PHE28 substituted by SER28, F28S) was found in Slc25a14, which showed the strongest differentiation signal (F_{ST} = 0.73) (Figures 4A and 4B). The three-dimensional structure of Slc25a14 based on homology modeling showed that the substitution (F28S) occurred in the loop region at the N terminal and increased its distance to the adjacent α-helix due to the changes in the electric charge and polarity of the amino acids (Figure 4C). To further evaluate the functional impact of the variant, we aligned the mutant Slc25a14 with its orthologs in diverse mammals and found that PHE28 is highly conserved in all the other animals we examined (Figure S17). The prediction of the functional effects of this variant supports that the F28S substitution is deleterious. All these results imply that F28S is likely the causal mutation for the Slc25a14 sweep in the high-altitude group.

To infer the genetic basis of adaptation shared by high-altitude adaptation and hibernation, we identified 116 DEGs in selective sweep regions. Among them, 25 DEGs were related to the HIF-1 pathway (25/62) and mainly belonged to the categories of DNA repair, angiogenesis, fatty acid metabolism, apoptosis/cell cycle, and heat generation (Table S23). A high proportion of selective sweep genes were differentially expressed during the torpor/arousal cycle, indicating a shared genetic influence between these two biological processes (Figure 4D). Among these genes, Bex1, Apln, kcne1l, Med12, Dad1, and Fgf16 were significantly upregulated in the liver or brain at the torpor stage (Figure 4D), suggesting their involvement in both hibernation and plateau adaptation. Previous studies have reported that these genes are involved in many physiological processes, including neuronal differentiation, liver regeneration, angiogenesis, and energy metabolism (Yu et al., 2016; Gu et al., 2018; Wysocka et al., 2018; Vilar et al., 2006; He et al., 2015; Rulifson et al., 2017). Further experimental validation is needed in future studies.

ψAamp Evolution and Potential Role in High-Altitude Adaptation

Pseudogenes are being increasingly recognized as an important regulatory factor in adaptive phenotypic diversification (Grander and Johnsson, 2016). Associated with selective sweep signals, we identified 27 candidate pseudogenes involved in extremely-high-altitude adaption, which represented 1.83% of the total genome-wide pseudogenes (Table S22). This ratio is almost comparable to that of functional genes with 1.78%, suggesting their non-negligible roles in extreme environmental adaptation. Consistent with the population genotype divergence found in the protein-coding genes, we also observed nearly fixed mutations in selective pseudogenes, such as ψAamp, ψAdl1, and ψRnf114, in the high-altitude group. Among

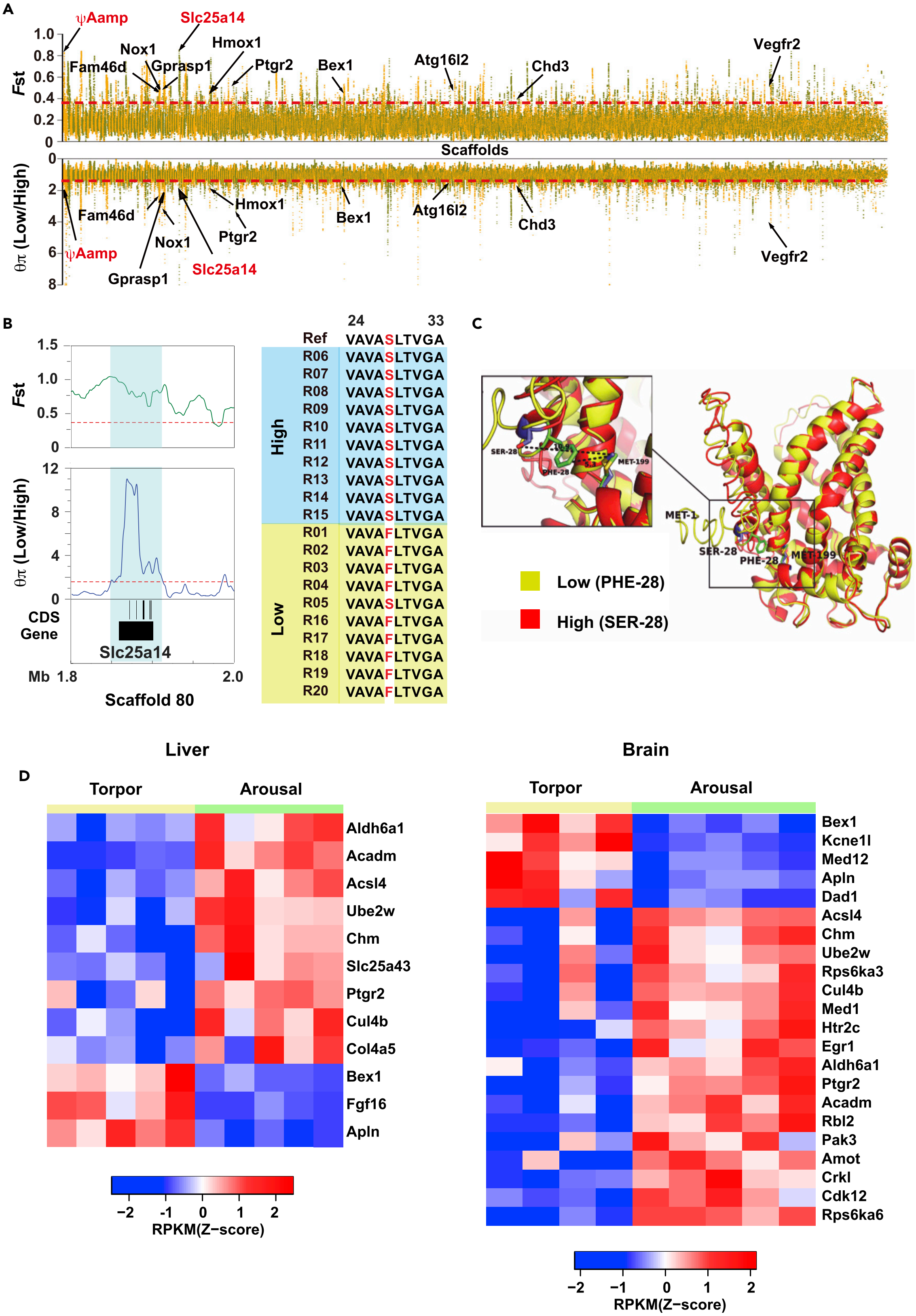

A
ψAamp
Slc25a14
1.0
0.8
0.6
0.4
0.2
0
Fst
Nox1
Hmox1
Fam46d
Gprasp1
Ptgr2
Bex1
Atg16l2
Chd3
Vegfr2
Scaffolds
0
2
4
6
8
θπ (Low/High)
Fam46d
Hmox1
Bex1
Atg16l2
Chd3
Nox1
Ptgr2
ψAamp
Slc25a14
Gprasp1
Vegfr2
B
1.5
1.0
0.5
0
Fst
12
8
4
0
θπ (Low/High)
CDS
Gene
Slc25a14
Mb 1.8
2.0
Scaffold 80
24
33
Ref VAVASLTVGA
High
R06 VAVASLTVGA
R07 VAVASLTVGA
R08 VAVASLTVGA
R09 VAVASLTVGA
R10 VAVASLTVGA
R11 VAVASLTVGA
R12 VAVASLTVGA
R13 VAVASLTVGA
R14 VAVASLTVGA
R15 VAVASLTVGA
Low
R01 VAVAFLTVGA
R02 VAVAFLTVGA
R03 VAVAFLTVGA
R04 VAVAFLTVGA
R05 VAVASLTVGA
R16 VAVAFLTVGA
R17 VAVAFLTVGA
R18 VAVAFLTVGA
R19 VAVAFLTVGA
R20 VAVAFLTVGA
C
SER-28
PHE-28
MET-199
MET-1
SER-28
PHE-28
MET-199
Low (PHE-28)
High (SER-28)
D
Liver
Torpor
Arousal
Aldh6a1
Acadm
Acsl4
Ube2w
Chm
Slc25a43
Ptgr2
Cul4b
Col4a5
Bex1
Fgf16
Apln
−2
-1
0
1
2
RPKM(Z−score)
Brain
Torpor
Arousal
Bex1
Kcne1l
Med12
Apln
Dad1
Acsl4
Chm
Ube2w
Rps6ka3
Cul4b
Med1
Htr2c
Egr1
Aldh6a1
Ptgr2
Acadm
Rbl2
Pak3
Amot
Crkl
Cdk12
Rps6ka6
−2
−1
0
1
2
RPKM(Z−score)

Figure 4. Signature of Selection in High-Altitude Himalayan Marmots and Overlap with Differentially Expressed Genes during Hibernation

(A) Manhattan plot of F_{st} and θ_π ratios between high- and low-altitude groups. F_{st} and θ_π ratios were calculated for each 100-Kb autosomal window. The red dashed lines denote the thresholds of F_{st} = 0.37 and θ_π ratios = 1.57. Ten genes and one pseudogene related to plateau adaptation are shown on their corresponding loci. Slc25a14 and ψAamp with a high selective signal are highlighted in red.

(B) F_{st} values (top) and θ_π ratios (middle) around the Slc25a14 locus. Genomic regions above the red dashed lines (top 5%) were defined as selective sweep regions and highlighted in light blue. F_{ST} and θ_π ratios were calculated for each 10-Kb window. Gene and its coding sequence (CDS) are shown in black box-and bars, respectively (bottom).

(C) The overlapped homology modeling of Slc25a14 from either high- or low-altitude Himalayan marmots. The enlarged view of the three-dimensional structure on the left shows a putative interaction between PHE-28 (low-altitude)/SER-28 (high-altitude) in a loop region at the N terminus and MET-199 in the α-helix. Distance measurements between pairs of atoms are marked with black dashed lines (low-altitude: 5.3 Å; high-altitude: 10.9 Å). PHE, phenylalanine; SER, serine; MET, methionine.

(D) Heatmap of hibernation-related differentially expressed genes under selective pressure. Gene expression is transformed into *Z* scores.

See also Figures S13–S17, and Tables S19, S20, S21, S22, and S23.

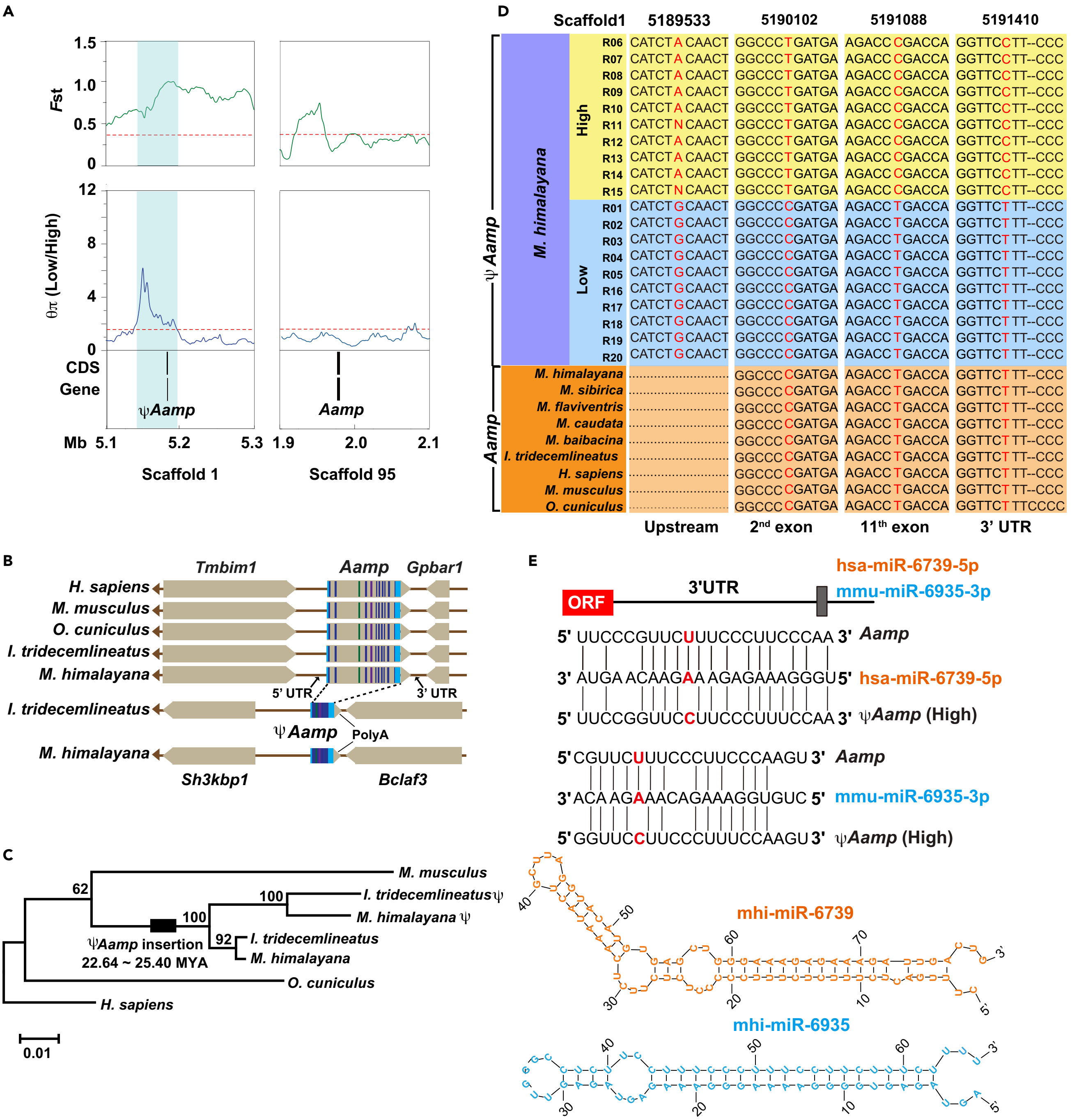
these pseudogenes, the angio-associated migratory cell protein pseudogene (ψAamp) possessed a complete gene structure and showed the highest F_{st} value (F_{st} = 0.67), but Aamp was not detected as a selective sweep gene (Figures 4A and 5A). The structure of ψAamp shows that it is an integrated processed pseudogene whose mRNA was inserted into the genome by retrotransposition (Figure 5B). It harbors premature stop codons (208th nucleotide, C > T), deletions/insertions, and frameshift mutations that abrogate its translation into a functional protein (Figure S18). The syntenic relationship of Aamp and ψAamp indicated that Aamp is highly conserved in different mammals. However, ψAamp is specifically found in ground squirrel and Himalayan marmot and preserved highly conservative syntenic locus (Figure 5B). Further phylogenetic analysis showed that ψAamp and Aamp belonged to two independent branches. Insertion time estimation based on sequence divergence suggested that pseudogene insertion events occurred at 22.64–25.40 MYA, before the split of Himalayan marmot and ground squirrel (Figure 5C). These findings support the hypothesis that this old pseudogene is specific to Sciuridae and was inherited from their common ancestor. However, Himalayan marmot showed a smaller sequence distance between Aamp and ψAamp compared with that of ground squirrel, suggesting the existence of selective pressure to maintain its genetic elements (Table S24).

As an important regulatory region, the 3′ UTR is a classically targeted locus of microRNA (miRNA) in animals (Yatsenko et al., 2014). Mutations in this region may influence mRNA abundance and phenotype divergence (Didiano and Hobert, 2008). We found four fixed substitutions in the upstream region, second exon, 11th exon, and 3′ UTR of ψAamp in the high-altitude group (Figure 5D). Within the 3′ UTR, we found a perfectly conserved seed match, which crossed the T > C substitution, for the ψAamp/Aamp-targeting hsa-miR-6739-5p and mmu-miR-6935-3p (Figure 5E). Pre-miRNA structure simulation showed that they hold stable hairpin structure in the Himalayan marmot genome (Figure 5E). Moreover, we detected transcription of ψAamp and Aamp in the liver of Himalayan marmots (Figure S19). The expression level of Aamp was decreased in the high-altitude group (Figure S20). These results support the hypothesis that mutation in the 3′ UTR of ψAamp may decrease the miRNA combination efficiency and thus influence the stability of Aamp in the high-altitude group.

DISCUSSION

In this study, we presented a draft genome of a male Himalayan marmot, which provides a valuable genomic resource to study its genome evolution and distinctive physiological features, including hypoxic adaptation and energy homeostasis. Phylogenetic analysis revealed that the formation of the Himalayan marmot species may be due to the uplift of the Tibetan Plateau. During the LGM, the Himalayan marmot was well adapted to the extreme environment.

A greater understanding of genetic diversity across the high- and low-altitude populations has enriched the knowledge of extreme environmental adaption. The strong signal of population differentiation and the apparent genotype difference found at the locus of Slc25a14 imply its important role in the adaptive responses of Himalayan marmots to hypoxia. Slc25a14 is neuroprotective through the regulation of mitochondrial function and oxidant production (Kwok et al., 2010; Kim-Han et al., 2001), and is also implicated in the maintenance of metabolic rate and adaptation thermoregulation (Yang et al., 2002; Yu et al., 2000). Interestingly, we found that ψAamp may be biologically active in mediating Aamp expression via competitive miRNA in the high-altitude group. Aamp is expressed in multiple cell types and mainly localized in the cytoplasm and membrane in vascular endothelial cells (Hu et al., 2016). The knockdown of Aamp impaired



Scaffold1		5189533	5190102	5191088	5191410
ψAamp, M. himalayana, High	R06	CATCTACAACT	GGCCCTGATGA	AGACCCGACCA	GGTTCCTT--CCC
	R07	CATCTACAACT	GGCCCTGATGA	AGACCCGACCA	GGTTCCTT--CCC
	R08	CATCTACAACT	GGCCCTGATGA	AGACCCGACCA	GGTTCCTT--CCC
	R09	CATCTACAACT	GGCCCTGATGA	AGACCCGACCA	GGTTCCTT--CCC
	R10	CATCTACAACT	GGCCCTGATGA	AGACCCGACCA	GGTTCCTT--CCC
	R11	CATCTNCAACT	GGCCCTGATGA	AGACCCGACCA	GGTTCCTT--CCC
	R12	CATCTACAACT	GGCCCTGATGA	AGACCCGACCA	GGTTCCTT--CCC
	R13	CATCTACAACT	GGCCCTGATGA	AGACCCGACCA	GGTTCCTT--CCC
	R14	CATCTACAACT	GGCCCTGATGA	AGACCCGACCA	GGTTCCTT--CCC
	R15	CATCTNCAACT	GGCCCTGATGA	AGACCCGACCA	GGTTCCTT--CCC
ψAamp, M. himalayana, Low	R01	CATCTGCAACT	GGCCCCGATGA	AGACCTGACCA	GGTTCTTT--CCC
	R02	CATCTGCAACT	GGCCCCGATGA	AGACCTGACCA	GGTTCTTT--CCC
	R03	CATCTGCAACT	GGCCCCGATGA	AGACCTGACCA	GGTTCTTT--CCC
	R04	CATCTGCAACT	GGCCCCGATGA	AGACCTGACCA	GGTTCTTT--CCC
	R05	CATCTGCAACT	GGCCCCGATGA	AGACCTGACCA	GGTTCTTT--CCC
	R16	CATCTGCAACT	GGCCCCGATGA	AGACCTGACCA	GGTTCTTT--CCC
	R17	CATCTGCAACT	GGCCCCGATGA	AGACCTGACCA	GGTTCTTT--CCC
	R18	CATCTGCAACT	GGCCCCGATGA	AGACCTGACCA	GGTTCTTT--CCC
	R19	CATCTGCAACT	GGCCCCGATGA	AGACCTGACCA	GGTTCTTT--CCC
	R20	CATCTGCAACT	GGCCCCGATGA	AGACCTGACCA	GGTTCTTT--CCC
Aamp	M. himalayana	....	GGCCCCGATGA	AGACCTGACCA	GGTTCTTT--CCC
	M. sibirica	....	GGCCCCGATGA	AGACCTGACCA	GGTTCTTT--CCC
	M. flaviventris	....	GGCCCCGATGA	AGACCTGACCA	GGTTCTTT--CCC
	M. caudata	....	GGCCCCGATGA	AGACCTGACCA	GGTTCTTT--CCC
	M. baibacina	....	GGCCCCGATGA	AGACCTGACCA	GGTTCTTT--CCC
	I. tridecemlineatus	....	GGCCCCGATGA	AGACCTGACCA	GGTTCTTT--CCC
	H. sapiens	....	GGCCCCGATGA	AGACCTGACCA	GGTTCTTT--CCC
	M. musculus	....	GGCCCCGATGA	AGACCTGACCA	GGTTCTTT--CCC
	O. cuniculus	....	GGCCCCGATGA	AGACCTGACCA	GGTTCTTTTCCCC
		Upstream	2nd exon	11th exon	3' UTR

Figure 5. ψAamp Evolution and Potential Role in High-Altitude Adaptation

(A) F_{st} values (top) and θ_{π} ratios (middle) around ψAamp and Aamp loci. Genomic regions above the red dashed lines (top 5%) were defined as selective sweep regions and highlighted in light blue. F_{st} and θ_{π} ratios were calculated for each 10-Kb window. Gene and coding sequence (CDS) are shown in black bars (bottom).

(B) The speculated formation of ψAamp. The collinearity of Aamp is shown among humans, the mouse, the rabbit, the ground squirrel, and the Himalayan marmot. In the Aamp gene box, the lines indicate exons. The mature mRNA of Aamp with intact polyA was presumably plugged back into the ancestral genome through retrotransposition and formed the new processed pseudogene, ψAamp. The collinearity of ψAamp found in the ground squirrel and Himalayan marmot is shown in the bottom.

(C) A phylogenetic analysis of Aamp and ψAamp. A putative insertion event of ancestral ψAamp was marked with a black box.

(D) Alignments of ψAamp and Aamp SNPs. SNP locations were numbered according to the ψAamp coordinate in the Himalayan marmot genome.

Figure 5. *Continued*

(E) Prediction of miRNAs binding to 3′ UTRs of Aamp and ψAamp (top). Aamp and ψAamp 3′ UTR contained the seed sequence of hsa-miR-6739-5p and mmu-miR-6935-3p. The 3′ UTR substitution (T > C in red) of ψAamp was located in the seed region. Prediction of the secondary structure of two pre-miRNAs in Himalayan marmot (bottom).

See also Figures S18–S20, and Tables S22, and S24.

VEGF-induced endothelial cell migration and angiogenesis (Hu et al., 2016). The downregulation of Aamp in the high-altitude group may be a protective strategy to prevent excess angiogenesis under extremely hypoxic conditions. Bearing in mind that Himalayan marmot is a hibernating animal, a close connection between plateau adaptation and hibernation was supported by a high proportion of selective sweep genes that were differentially expressed during the torpor/arousal cycle. Meanwhile, complement and coagulation cascades and pluripotency of stem cell signaling pathways may be implicated in the protective strategy of the brain for cold resistance. The comprehensive characterization of the Himalayan marmot genome along with whole-genome resequencing data of additional marmots presented herein provide a broad view for elucidating its evolutionary events and environmental adaptation. The identification of distinctive genetic traits will contribute to its potential medical applications.

Limitations of the Study

Here, we provide a comprehensive framework for understanding the genetic adaptation to the harsh environment in Himalayan marmot. Although the quality and contiguity of the Himalayan marmot genome assembly is carefully validated and generally reliable for the current study, the contig/scaffold N50 does not reach the best level of the mammalian genome assemblies. With the development of high-throughput sequencing technology, the assembly quality will be improved in future work. In addition, functional experimental assays should be performed to further validate the selected sweep genes and identify the targets involved in hypoxic and cold adaptation.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Transparent Methods, 20 figures, and 24 tables and can be found with this article online at https://doi.org/10.1016/j.isci.2018.11.034.

ACKNOWLEDGMENTS

This work was supported by grants from the Natural Science Foundation of Shaanxi Province (2014PT013 to E.L., 2017JZ028 to J.F.), National Natural Science Foundation of China (81770457 to J.F., 81370379 to S.Z., 31160436 to Y.T.), Qinghai Province Grant for Science and Technology Capacity Building (2009-TY-05 to Z.W.), and China Postdoctoral Science Foundation Grant (2016M590954 to L.B.). We thank Professor Jun Yu (Beijing Institute of Genomics, Chinese Academy of Sciences) for his helpful suggestions.

AUTHOR CONTRIBUTIONS

E.L., J.F., and H.Z. conceived the project. E.L., J.F., L.B., B.L., C.J., S.Z., and D.L. designed the experiments. C.J., B.L., S.L., P.Y., X.L., X.F., F.H., M.L., and H.Y. performed genome sequencing and bioinformatics analyses. L.B., S.Z., R.W., W.W., N.G., and H.L. conducted bench experiments. Y.T., L.B., Z.W., S.Y., G.F., M.T.B., and J.Z. provided samples and specimen dissections. L.B., B.L., C.J., P.Y., X.L., and X.F. prepared figures and tables. L.B., B.L., C.J., and S.Z. wrote the manuscript. E.L. and J.F. revised the manuscript. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 6, 2018
Revised: November 19, 2018
Accepted: November 20, 2018
Published: December 20, 2018

REFERENCES

Cardini, A. (2003). The geometry of the marmot (Rodentia: Sciuridae) mandible: phylogeny and patterns of morphological evolution. Syst. Biol. *52*, 186–205.

Cardini, A., and O'Higgins, P. (2005). Post-natal ontogeny of the mandible and ventral cranium in Marmota species (Rodentia, Sciuridae): allometry and phylogeny. Zoomorphology *124*, 189–203.

Didiano, D., and Hobert, O. (2008). Molecular architecture of a miRNA-regulated 3' UTR. RNA *14*, 1297–1317.

Fedorov, V.B., Goropashnaya, A.V., Toien, O., Stewart, N.C., Chang, C., Wang, H., Yan, J., Showe, L.C., Showe, M.K., and Barnes, B.M. (2011). Modulation of gene expression in heart and liver of hibernating black bears (*Ursus americanus*). BMC Genomics *12*, 171.

Ge, R.L., Cai, Q., Shen, Y.Y., San, A., Ma, L., Zhang, Y., Yi, X., Chen, Y., Yang, L., Huang, Y., et al. (2013). Draft genome sequence of the Tibetan antelope. Nat. Commun. *4*, 1858.

Geiser, F. (2013). Hibernation. Curr. Biol. *23*, R188–R193.

Gou, X., Wang, Z., Li, N., Qiu, F., Xu, Z., Yan, D., Yang, S., Jia, J., Kong, X., Wei, Z., et al. (2014). Whole-genome sequencing of six dog breeds from continuous altitudes reveals adaptation to high-altitude hypoxia. Genome Res. *24*, 1308–1315.

Grander, D., and Johnsson, P. (2016). Pseudogene-expressed RNAs: emerging roles in gene regulation and disease. Curr. Top. Microbiol. Immunol. *394*, 111–126.

Gu, Y., Wei, W., Cheng, Y., Wan, B., Ding, X., Wang, H., Zhang, Y., and Jin, M. (2018). A pivotal role of BEX1 in liver progenitor cell expansion in mice. Stem Cell Res. Ther. *9*, 164.

He, L., Xu, J., Chen, L., and Li, L. (2015). Apelin/APJ signaling in hypoxia-related diseases. Clin. Chim. Acta *451* (Pt B), 191–198.

Hoffmann, R.S., and Nadler, C.F. (1968). Chromosomes and systematics of some North American species of genus Marmota (Rodentia - Sciuridae). Experientia *24*, 740–742.

Hu, J.J., Qiu, J.H., Zheng, Y.M., Zhang, T., Yin, T.Y., Xie, X., and Wang, G.X. (2016). AAMP regulates endothelial cell migration and angiogenesis through RhoA/Rho kinase signaling. Ann. Biomed. Eng. *44*, 830–832.

Kim-Han, J.S., Reichert, S.A., Quick, K.L., and Dugan, L.L. (2001). BMCP1: a mitochondrial uncoupling protein in neurons which regulates mitochondrial function and oxidant production. J. Neurochem. *79*, 658–668.

Kwok, K.H., Ho, P.W., Chu, A.C., Ho, J.W., Liu, H.F., Yiu, D.C., Chan, K.H., Kung, M.H., Ramsden, D.B., and Ho, S.L. (2010). Mitochondrial UCP5 is neuroprotective by preserving mitochondrial membrane potential, ATP levels, and reducing oxidative stress in MPP+ and dopamine toxicity. Free Radic. Biol. Med. *49*, 1023–1035.

Laursen, W.J., Mastrotto, M., Pesta, D., Funk, O.H., Goodman, J.B., Merriman, D.K., Ingolia, N., Shulman, G.I., Bagriantsev, S.N., and Gracheva, E.O. (2015). Neuronal UCP1 expression suggests a mechanism for local thermogenesis during hibernation. Proc. Natl. Acad. Sci. U S A *112*, 1607–1612.

Li, R.Q., Fan, W., Tian, G., Zhu, H.M., He, L., Cai, J., Huang, Q.F., Cai, Q.L., Li, B., Bai, Y.Q., et al. (2010). The sequence and de novo assembly of the giant panda genome. Nature *463*, 311–317.

Matthews, L. (1971). The Life of Mammals, in the Life of Mammals (Weinfield and Nicolson).

Nikol'skii, A.A., and Ulak, A. (2006). Key factors determining the ecological niche of the Himalayan marmot, *Marmota himalayana* Hodgson (1841). Russ. J. Ecol. *37*, 46–52.

Ou, J., Ball, J.M., Luan, Y., Zhao, T., Miyagishima, K.J., Xu, Y., Zhou, H., Chen, J., Merriman, D.K., Xie, Z., et al. (2018). iPSCs from a hibernator provide a platform for studying cold adaptation and its potential medical applications. Cell *173*, 851–863.e16.

Qiu, Q., Wang, L.Z., Wang, K., Yang, Y.Z., Ma, T., Wang, Z.F., Zhang, X., Ni, Z.Q., Hou, F.J., Long, R.J., et al. (2015). Yak whole-genome resequencing reveals domestication signatures and prehistoric population expansions. Nat. Commun. *6*, 10283.

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. Nat. Genet. *44*, 946–949.

Rulifson, I.C., Collins, P., Miao, L., Nojima, D., Lee, K.J., Hardy, M., Gupte, J., Hensley, K., Samayoa, K., Cam, C., et al. (2017). In vitro and in vivo analyses reveal profound effects of fibroblast growth factor 16 as a metabolic regulator. J. Biol. Chem. *292*, 1951–1969.

Seim, I., Fang, X., Xiong, Z., Lobanov, A.V., Huang, Z., Ma, S., Feng, Y., Turanov, A.A., Zhu, Y., Lenz, T.L., et al. (2013). Genome analysis reveals insights into physiology and longevity of the Brandt's bat *Myotis brandtii*. Nat. Commun. *4*, 2212.

Shrestha, T. (2016). Marmota himalayana. The IUCN Red List of Threatened Species, e.T12826A115106426. http://dx.doi.org/10.2305/IUCN.UK.2016-3.RLTS.T12826A22258911.en.

Smith, A.T., Xie, Y., Hoffmann, R.S., Mackinnon, J., Wilson, D.E., and Wozencraft, W.C. (2010). A Guide to the Mammals of China (Princeton University Press).

Steppan, S.J., Akhverdyan, M.R., Lyapunova, E.A., Fraser, D.G., Vorontsov, N.N., Hoffmann, R.S., and Braun, M.J. (1999). Molecular phylogeny of the marmots (Rodentia: Sciuridae): Tests of evolutionary and biogeographic hypotheses. Syst. Biol. *48*, 715–734.

Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell *126*, 663–676.

Thomas, W.K., and Martin, S.L. (1993). A recent origin of marmots. Mol. Phylogenet. Evol. *2*, 330–336.

Vilar, M., Murillo-Carretero, M., Mira, H., Magnusson, K., Besset, V., and Ibanez, C.F. (2006). Bex1, a novel interactor of the p75 neurotrophin receptor, links neurotrophin signaling to the cell cycle. EMBO J. *25*, 1219–1230.

Williams, D.R., Epperson, L.E., Li, W., Hughes, M.A., Taylor, R., Rogers, J., Martin, S.L., Cossins, A.R., and Gracey, A.Y. (2005). Seasonally hibernating phenotype assessed through transcript screening. Physiol. Genomics *24*, 13–22.

Wu, T. (2001). The Qinghai-Tibetan plateau: how high do tibetans live? High Alt. Med. Biol. *2*, 489–499.

Wysocka, M.B., Pietraszek-Gremplewicz, K., and Nowak, D. (2018). The role of apelin in cardiovascular diseases, obesity and cancer. Front. Physiol. *9*, 557.

Yang, X., Pratley, R.E., Tokraks, S., Tataranni, P.A., and Permana, P.A. (2002). UCP5/BMCP1 transcript isoforms in human skeletal muscle: relationship of the short-insert isoform with lipid oxidation and resting metabolic rates. Mol. Genet. Metab. *75*, 369–373.

Yatsenko, A.S., Marrone, A.K., and Shcherbata, H.R. (2014). miRNA-based buffering of the cobblestone-lissencephaly-associated extracellular matrix receptor dystroglycan via its alternative 3'-UTR. Nat. Commun. *5*, 4906.

Yu, W., Huang, X., Tian, X., Zhang, H., He, L., Wang, Y., Nie, Y., Hu, S., Lin, Z., Zhou, B., et al. (2016). GATA4 regulates Fgf16 to promote heart repair after injury. Development *143*, 936–949.

Yu, X.X., Mao, W., Zhong, A., Schow, P., Brush, J., Sherwood, S.W., Adams, S.H., and Pan, G. (2000). Characterization of novel UCP5/BMCP1 isoforms and differential regulation of UCP4 and UCP5 expression through dietary or temperature manipulation. FASEB J. *14*, 1611–1618.

Zhao, S.C., Zheng, P.P., Dong, S.S., Zhan, X.J., Wu, Q., Guo, X.S., Hu, Y.B., He, W.M., Zhang, S.N., Fan, W., et al. (2013). Whole-genome sequencing of giant pandas provides insights into demographic history and local adaptation. Nat. Genet. *45*, 67–U99.

Zheng, B.X., Xu, Q.Q., and Shen, Y.P. (2002). The relationship between climate change and Quaternary glacial cycles on the Qinghai-Tibetan Plateau: review and speculation. Quaternary. Int. *97-8*, 93–101.

Zhou, X.M., Wang, B.S., Pan, Q., Zhang, J.B., Kumar, S., Sun, X.Q., Liu, Z.J., Pan, H.J., Lin, Y., Liu, G.J., et al. (2014). Whole-genome sequencing of the snub-nosed monkey provides insights into folivory and evolutionary history. Nat. Genet. *46*, 1303–1310.

ISCI, Volume *11*

Supplemental Information

Hypoxic and Cold Adaptation Insights

from the Himalayan Marmot Genome

Liang Bai, Baoning Liu, Changmian Ji, Sihai Zhao, Siyu Liu, Rong Wang, Weirong Wang, Pu Yao, Xuming Li, Xiaojun Fu, Haiyan Yu, Min Liu, Fengming Han, Ning Guan, Hui Liu, Dongyuan Liu, Yuanqing Tao, Zhongdong Wang, Shunsheng Yan, Greg Florant, Michael T. Butcher, Jifeng Zhang, Hongkun Zheng, Jianglin Fan,Enqi Liu

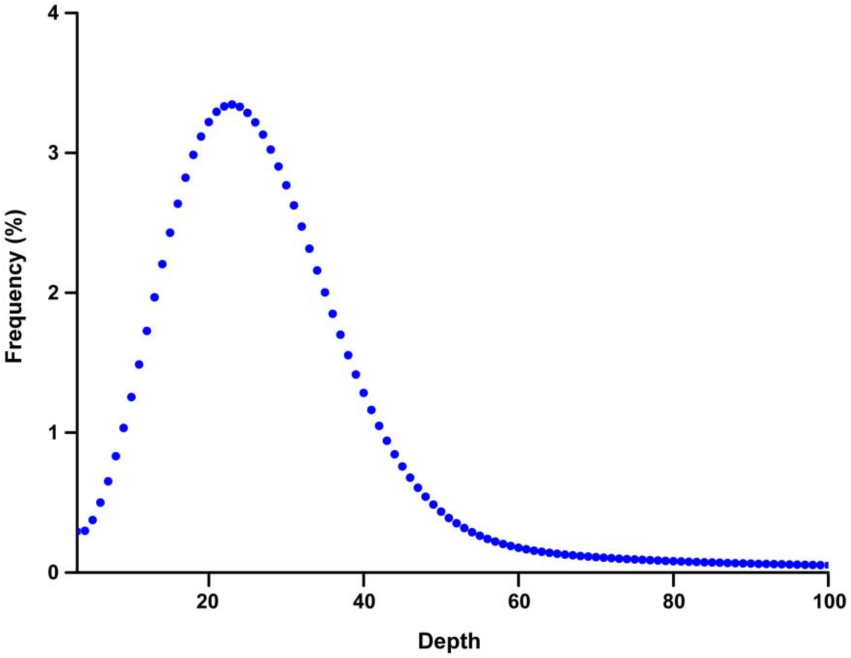


Figure S1. Distribution of 21-mer frequency, Related to Figure 1.

The frequency of each 21-mer in raw sequencing reads was calculated. The genome size of the Himalayan marmot was estimated to be ~2.33 Gb based on reads from short insert size libraries.

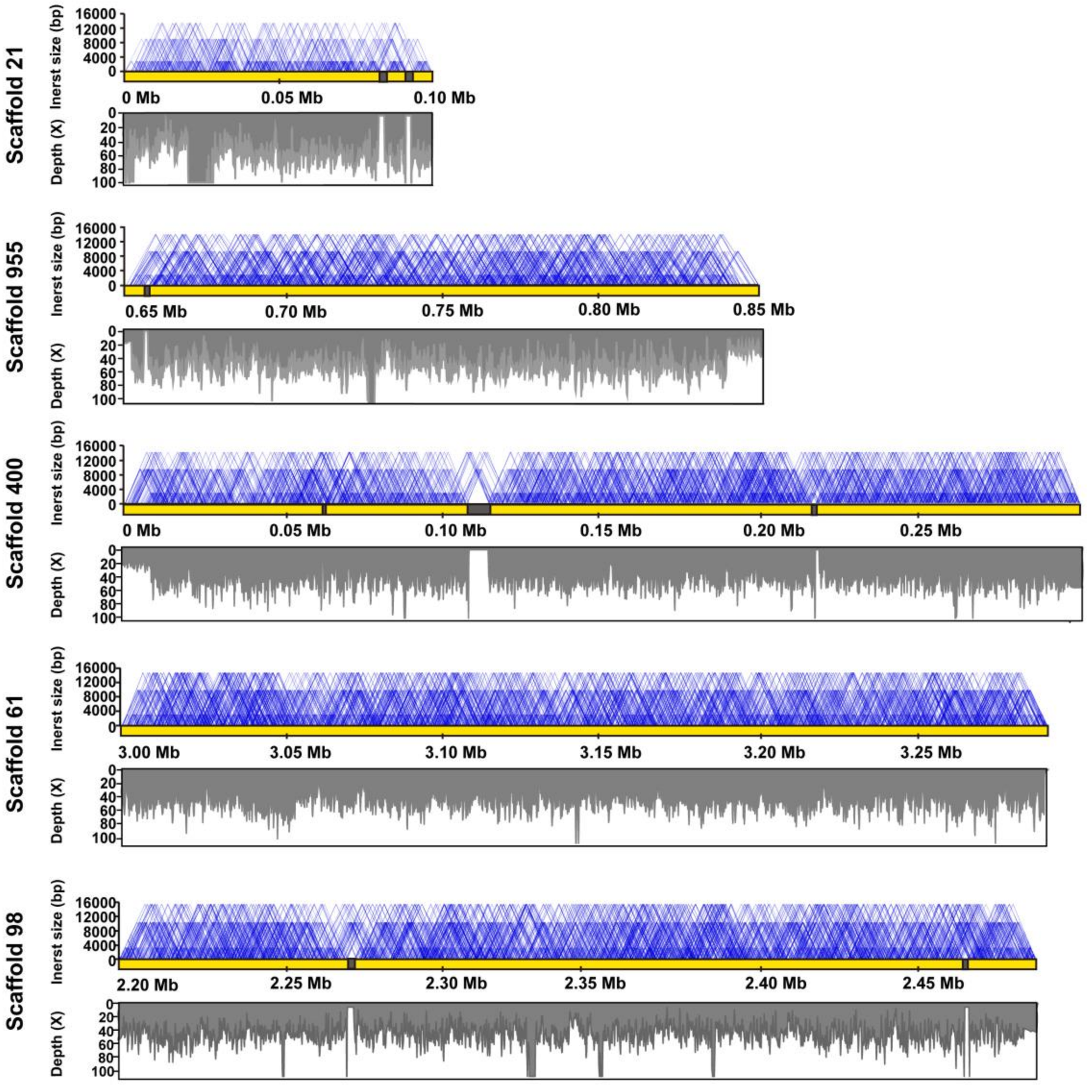


Figure S2. Mate-pair read relationships and depth distributions from three libraries (4 Kb, 10 Kb, and 15 Kb), , Related to Figure 1.

Five scaffolds (Scaffold 21, 955, 400, 61, and 98) were randomly selected to validate the assembly accuracy of the genome. High overall depth coverage and a corrected paired-end relationship indicate the high accuracy of genome assembly.

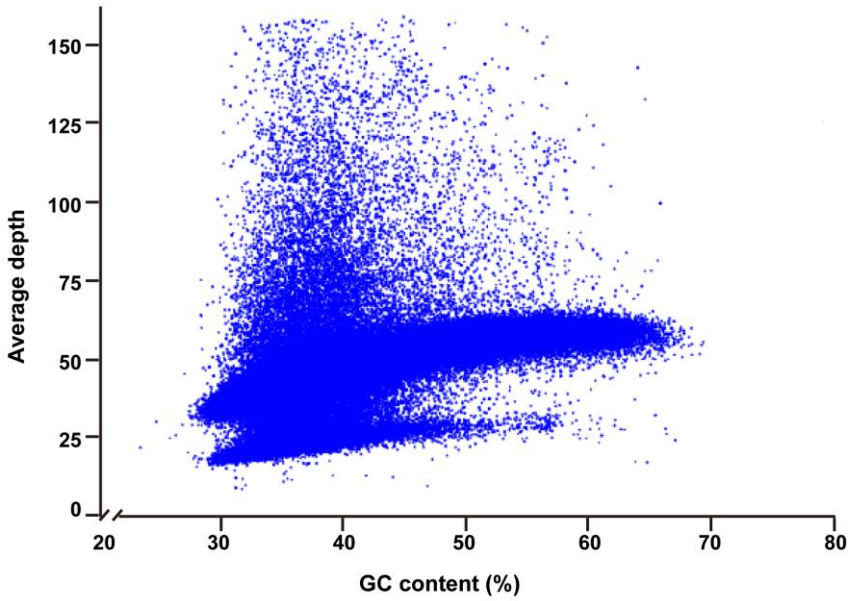


Figure S3. Depth and GC content of the Himalayan marmot genome, Related to Figure 1.

The depth of sex chromosome sequencing is 50% of the sequence depth of autosomes, and, thus, there is a low-depth fraction of the genome.

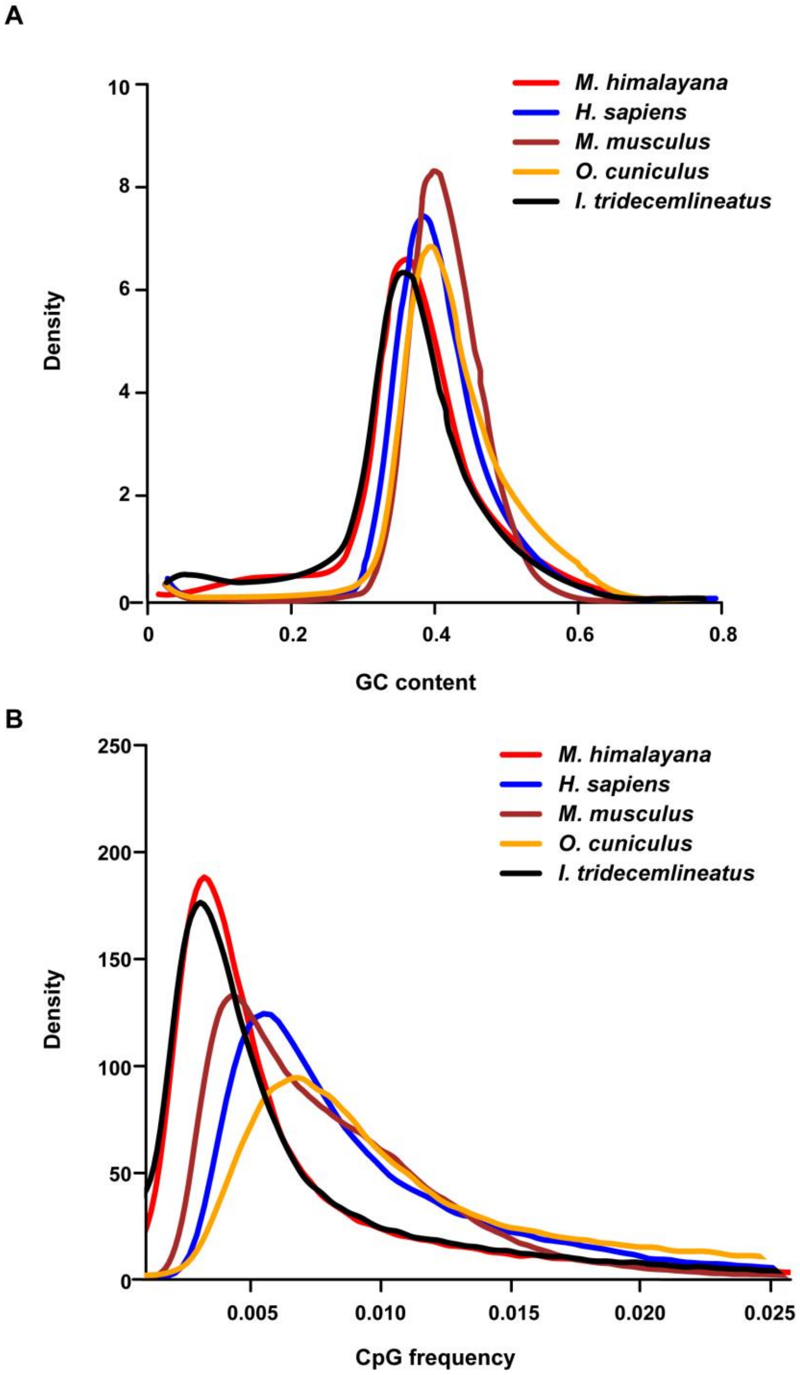


Figure S4. Comparison of GC content (A) and CpG frequency (B) in five mammals, Related to Figure 1.

The Himalayan marmot and ground squirrel, which belong to *Sciuridae*, share more similar profiles than other mammals. *M. himalayana*, Himalayan marmot; *H. sapiens*, human; *M. musculus*, mouse; *O. cuniculus*, rabbit; *I. tridecemlineatus*, ground squirrel.

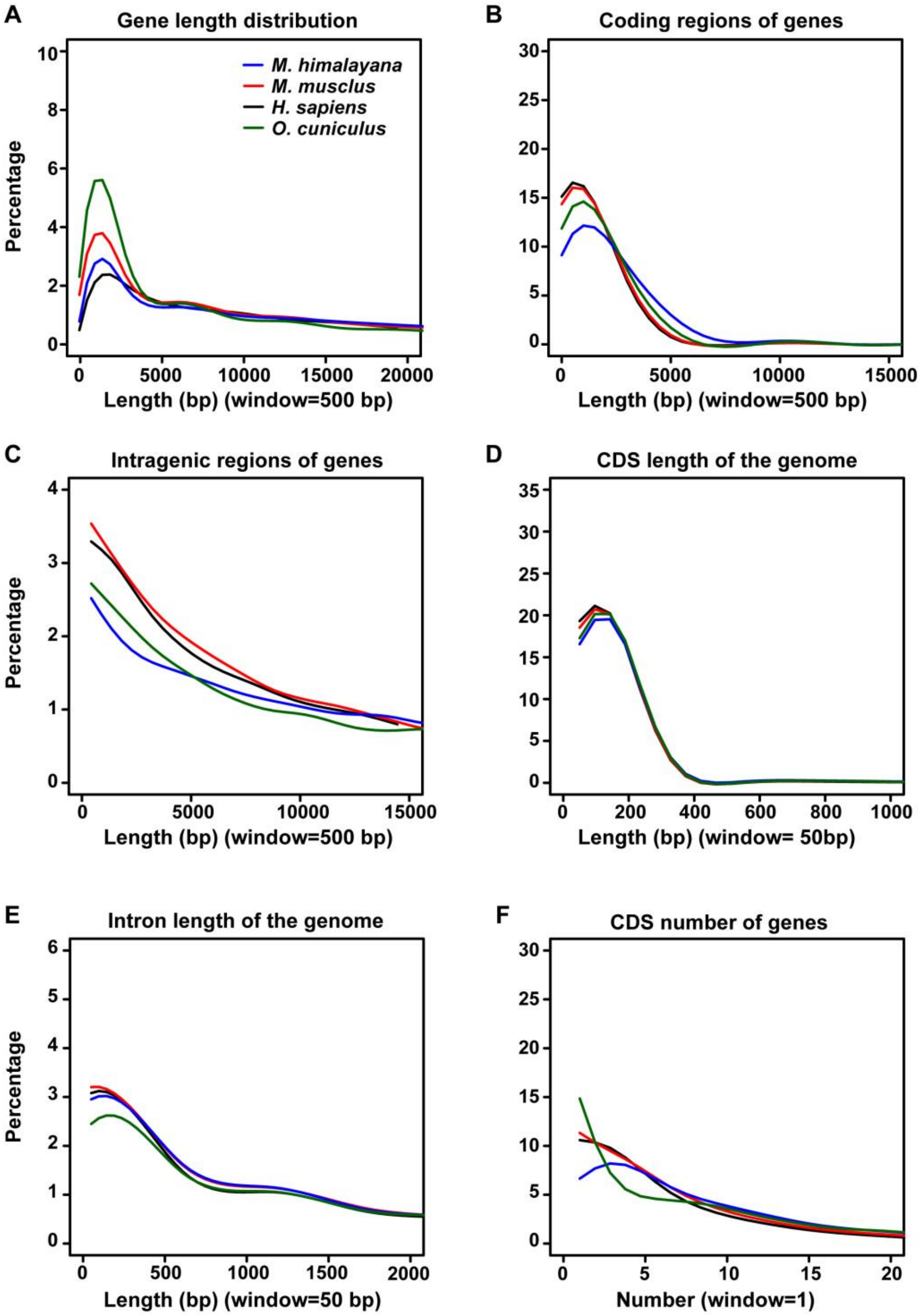


Figure S5. Comparison of gene features among four sequenced mammalian genomes, Related to Figure 1.

The gene features of Himalayan marmot are similar to these mammals with respect to all of the key parameters, indicating the high quality of gene structure annotation in the Himalayan marmot genome. *M. himalayana*, Himalayan marmot; *M. musculus*, mouse; *H. sapiens*, human; *O. cuniculus*, rabbit.

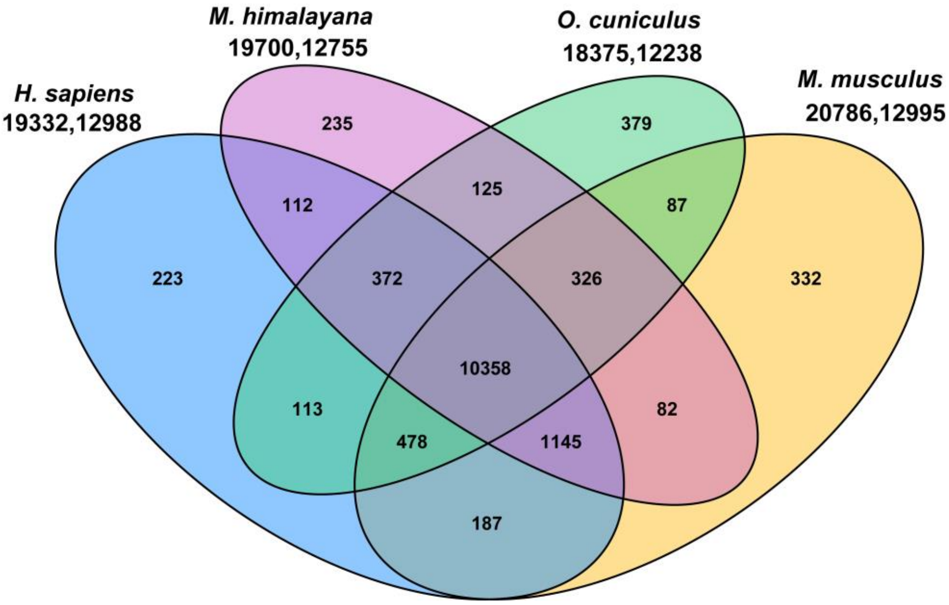


Figure S6. Venn plots showing specific gene families of Himalayan marmot, Related to Figure 1.
The digits below the species names indicate the number of clustered genes and families, respectively. *M. himalayana*, Himalayan marmot; *M. musculus*, mouse; *H. sapiens*, human; *O. cuniculus*, rabbit.

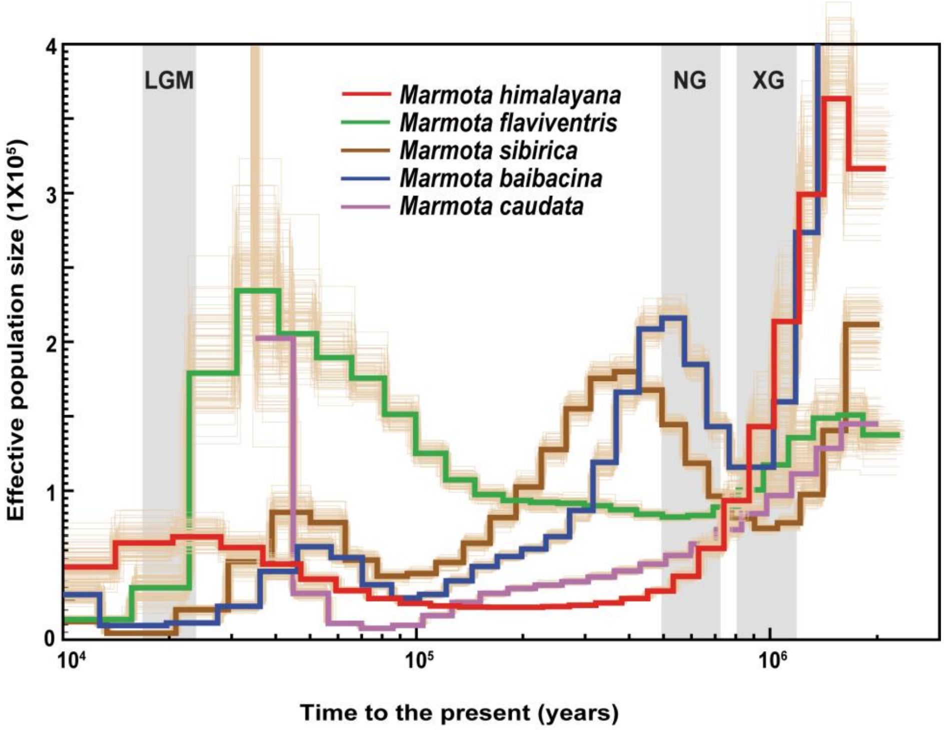


Figure S7. Demographic histories of five *Marmota* species reconstructed based on the pairwise sequentially Markovian coalescent model, Related to Figure 2. The periods of the last glacial maximum (LGM, ~20 kya), Naynayxungla Glaciation (NG, 0.5~0.78 mya), and Xixiabangma Glaciation (XG, 0.8~0.17 mya) are shaded in gray. *Marmota himalayana*, Himalayan marmot; *Marmota flaviventris*, yellow-bellied marmot; *Marmota sibirica*, Mongolian marmot; *Marmota baibacina*, gray marmot, *Marmota caudate,* long-tailed marmot.

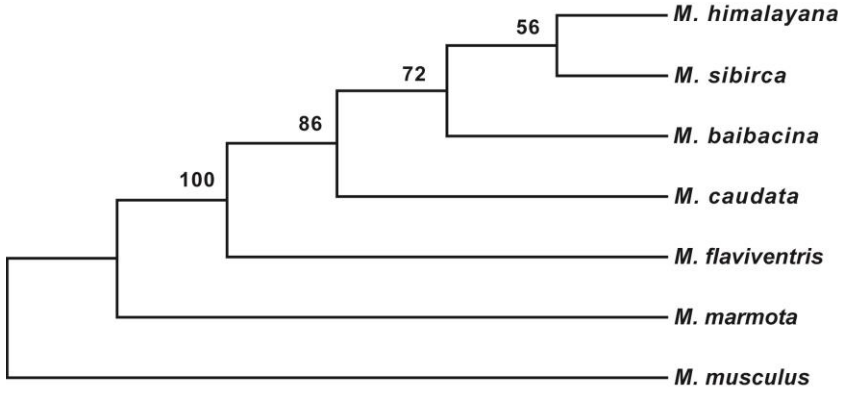


Figure S8. Phylogenetic tree of six *Marmota* species based on Y-chromosome gene sequence, Related to Figure 2.

Numbers along the nodes indicate the degree of the bootstrap-based support with 1000 replications. *M. himalayana*, Himalayan marmot; *M. sibirica*, Mongolian marmot; *M. baibacina*, gray marmot; *M. caudate,* long-tailed marmot; *M. flaviventris*, yellow-bellied marmot; *M. marmota*, Alpine marmot; *M. musculus*, mouse.

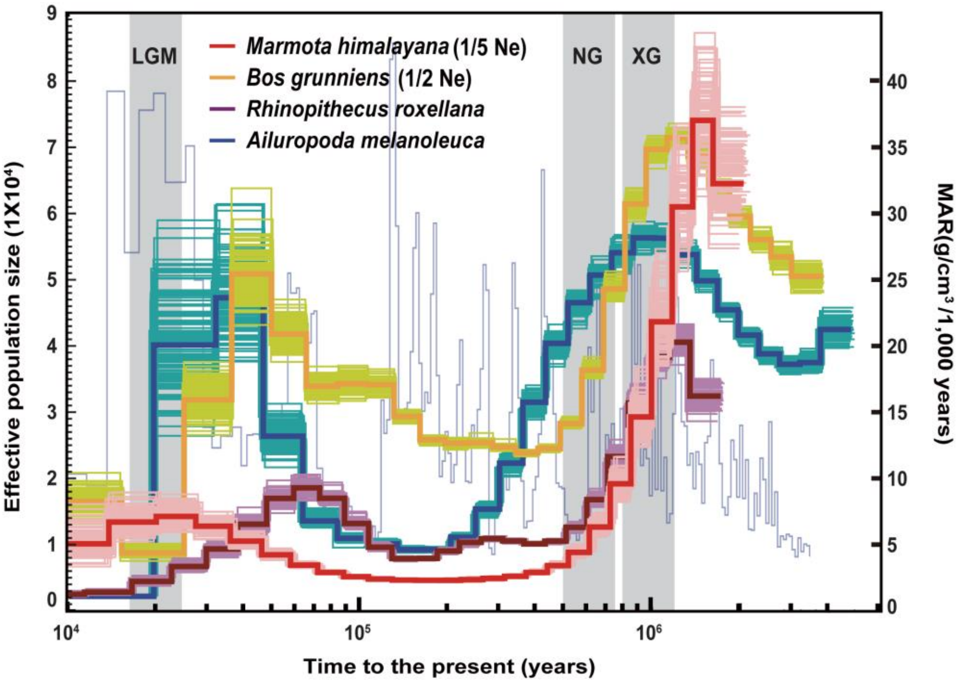


Figure S9. Demographic histories of Himalayan marmot reconstructed using pairwise sequentially Markovian coalescent model, Related to Figure 2.
The effective population size (*Ne*) of the Himalayan marmot and yak are rescaled to 1/5 and 1/2 of the actual *Ne*, respectively. The light blue lines show the mass accumulation rate (MAR) of the Chinese loess. The periods of the last glacial maximum (LGM, ~20 MYA), Naynayxungla Glaciation (NG, 0.5~0.78 MYA), and Xixiabangma Glaciation (XG, 0.8~0.17 MYA) are shaded in gray. g (generation time) = 3 years; μ (neutral mutation rate per generation) = 2.54 X 10^{-9}. *Marmota himalayana*, Himalayan marmot; *Bos grunniens,* yak; *Rhinopithecus roxellana*, golden snub-nosed monkey; *Ailuropoda melanoleuca*, giant panda.

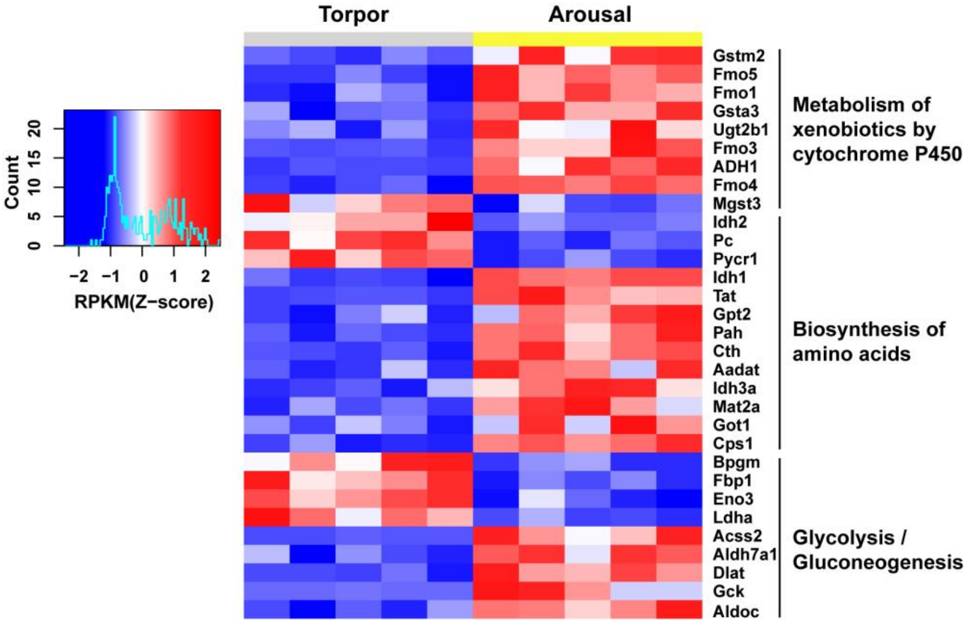


Figure S10. Differentially expressed genes in the liver of Himalayan marmot during the torpor/arousal cycle, Related to Figure 3.
Red and blue colors represent high and low gene expression, respectively. Gene expression was transformed into Z-scores and the numbers of gene counts are shown as blue lines in the color keys.

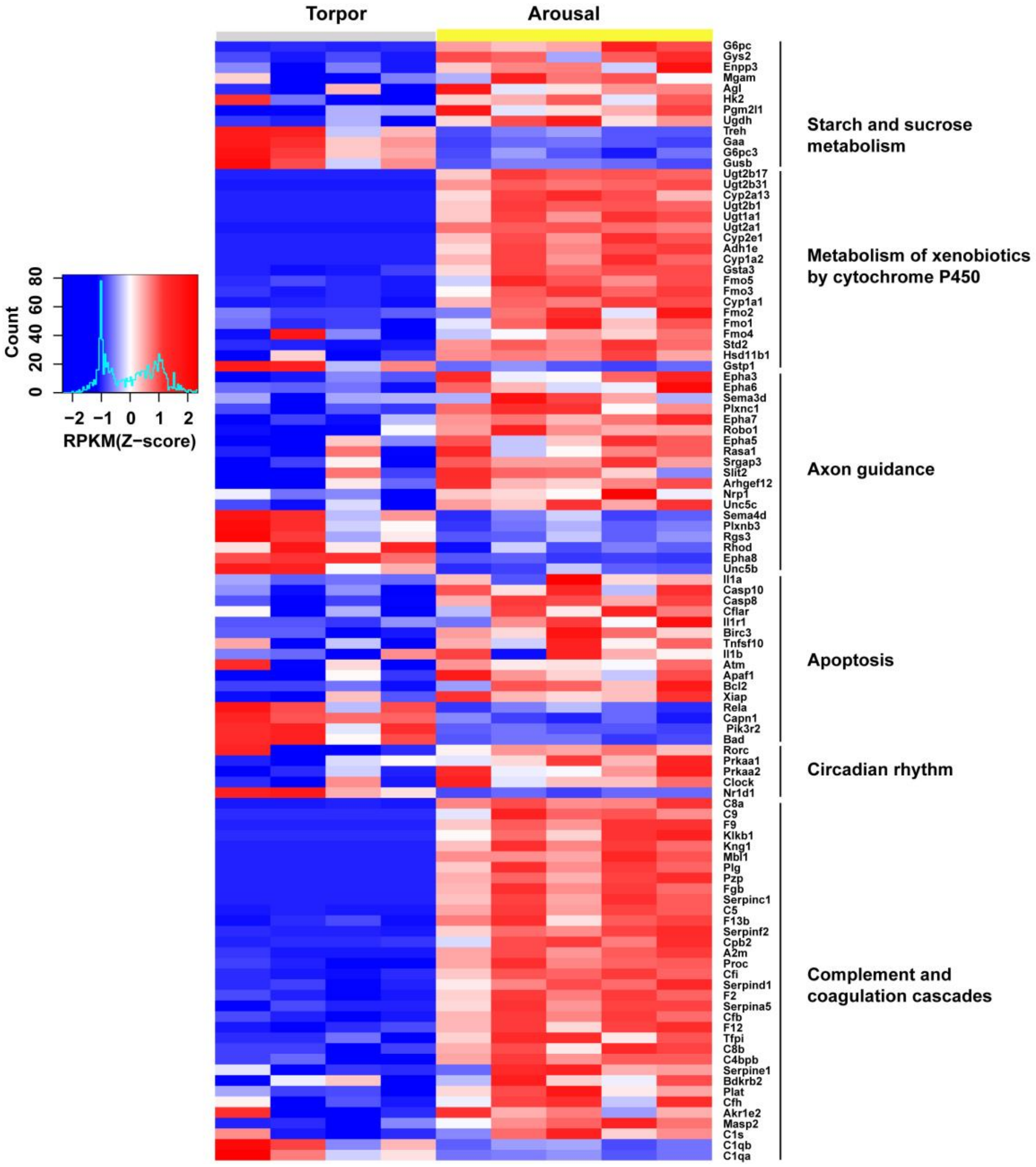


Figure S11. Differentially expressed genes in the brain of Himalayan marmot during torpor/arousal cycle, Related to Figure 3.

Red and blue colors represent high and low gene expression, respectively. Gene expression was transformed into Z-scores and the numbers of gene counts are shown as blue lines in the color keys.

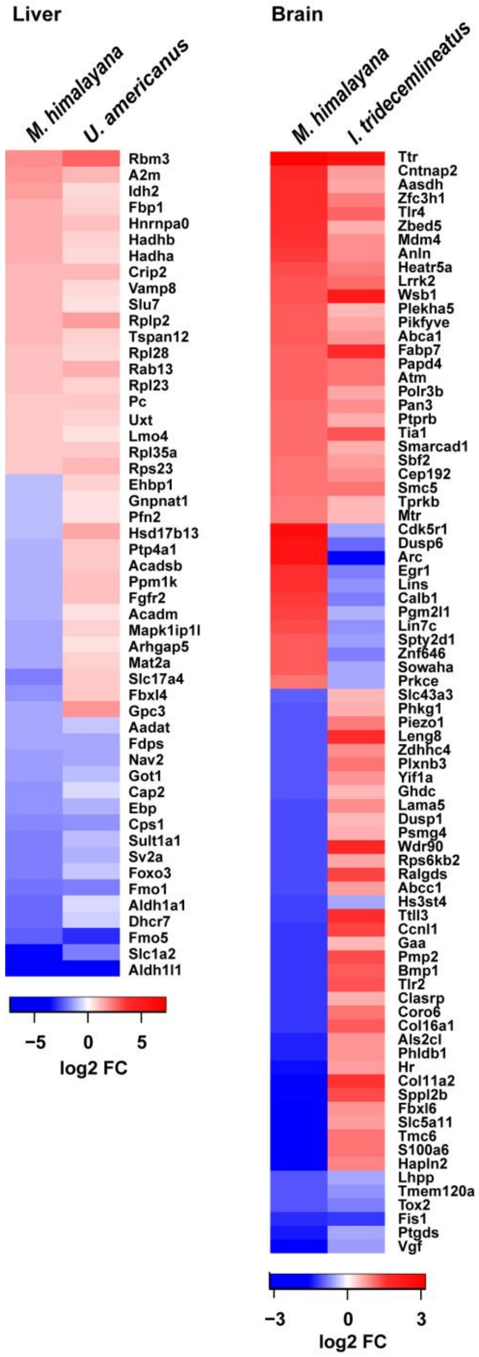


Figure S12. Comparison of differentially expressed genes in liver and brain between Himalayan marmot and black bear/ground squirrel, Related to Figure 3. Red and blue indicate an increase and decrease relative to the arousal stage, respectively. *M. himalayana*, Himalayan marmot; *U. americanus*, black bear; *I. tridecemlineatus*, ground squirrel. FC, fold change.

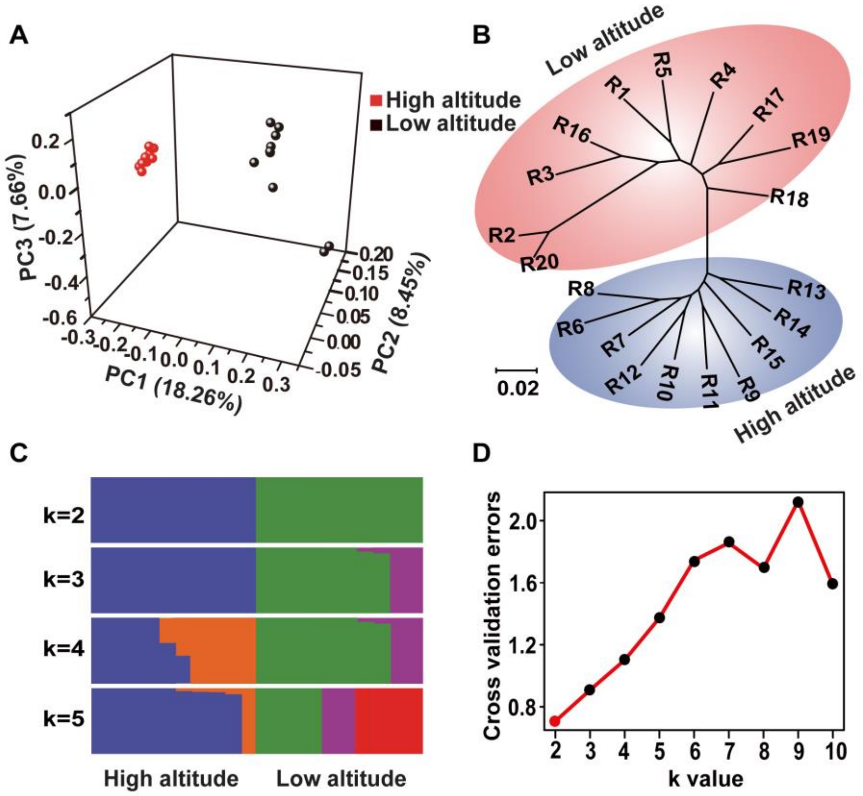


Figure S13. Genetic relationships and population structures of 20 Himalayan marmots based on all autosomal SNPs, Related to Figure 4.

(A) Three-way principal component analysis (PCA) plot. Percentages indicate the proportion of variance explained by each component.

(B) Unrooted maximum likelihood tree.

(C) Population structures with the number of ancestral clusters K from 2 to 5. Each color represents one ancestral cluster. The length of colored segments represents corresponding ancestry attributions.

(D) Cross validation errors of ancestral clusters K.

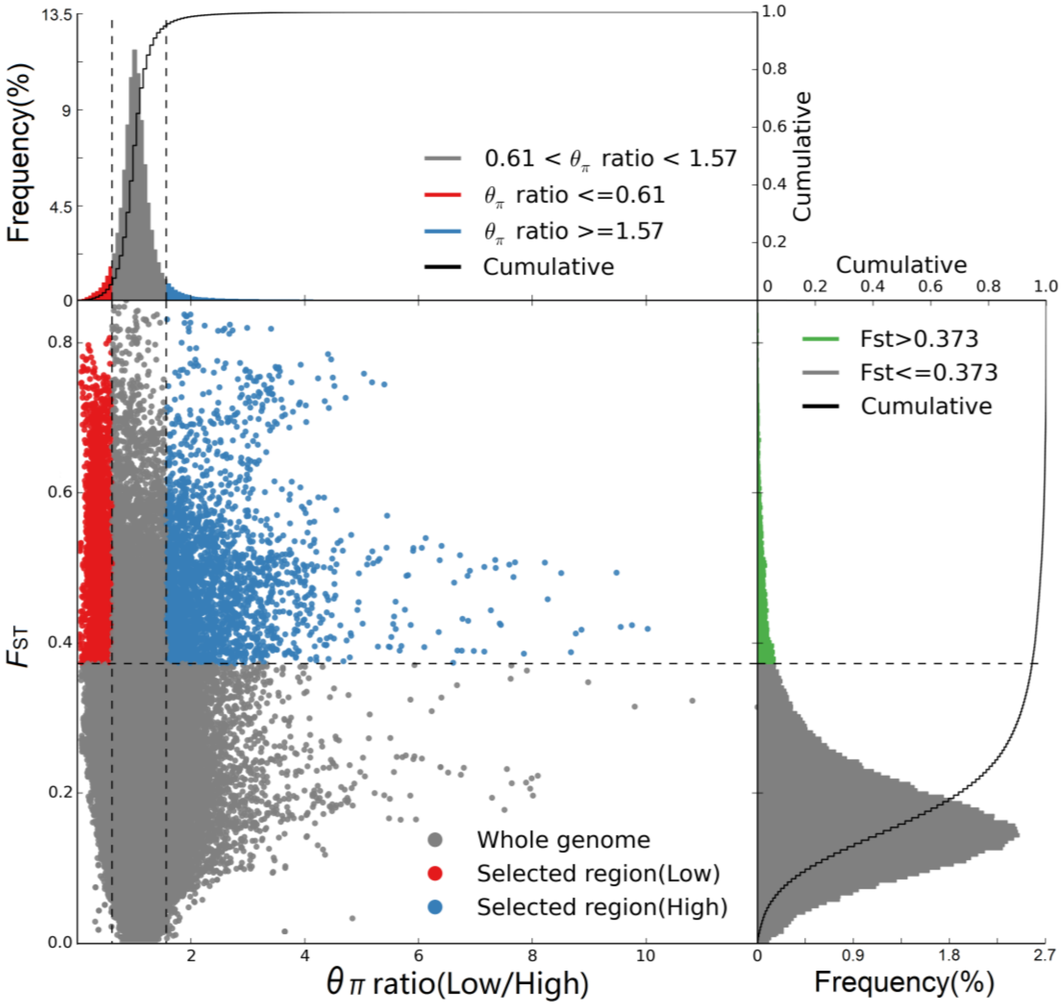


Figure S14. Distribution of F_{ST} values and θ_π ratios (θ_π low/θ_π high), Related to Figure 4.

Data points located to the right of the second vertical dashed line (corresponding to the 5% right tail of the empirical θ_π ratio distribution, where the θ_π ratio is 1.57), and above the horizontal dashed line (the 5% right tails of the empirical F_{ST} distribution, where F_{ST} is 0.373), are defined as selective sweep regions for the high-altitude group. Calculations were performed in 100-kb windows sliding in 10-kb steps.

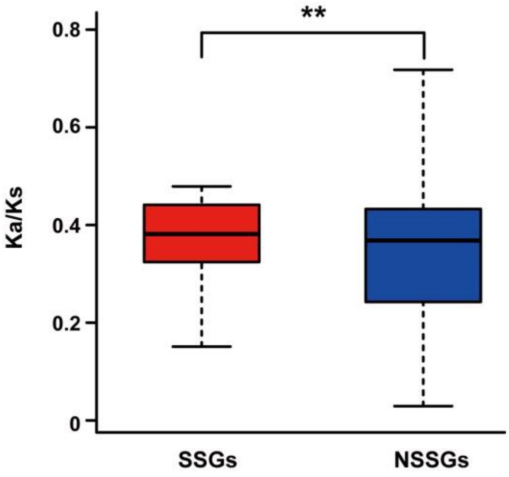


Figure S15. Ka/Ks values in SSGs and NSSGs, Related to Figure 4.
Selective sweep genes (SSGs) show a significantly faster evolutionary rate than nonselective sweep genes (NSSGs). (*p*<0.01, the Student's *t*-test; n=383 for SSGs and n= 21,226 for NSSGs).

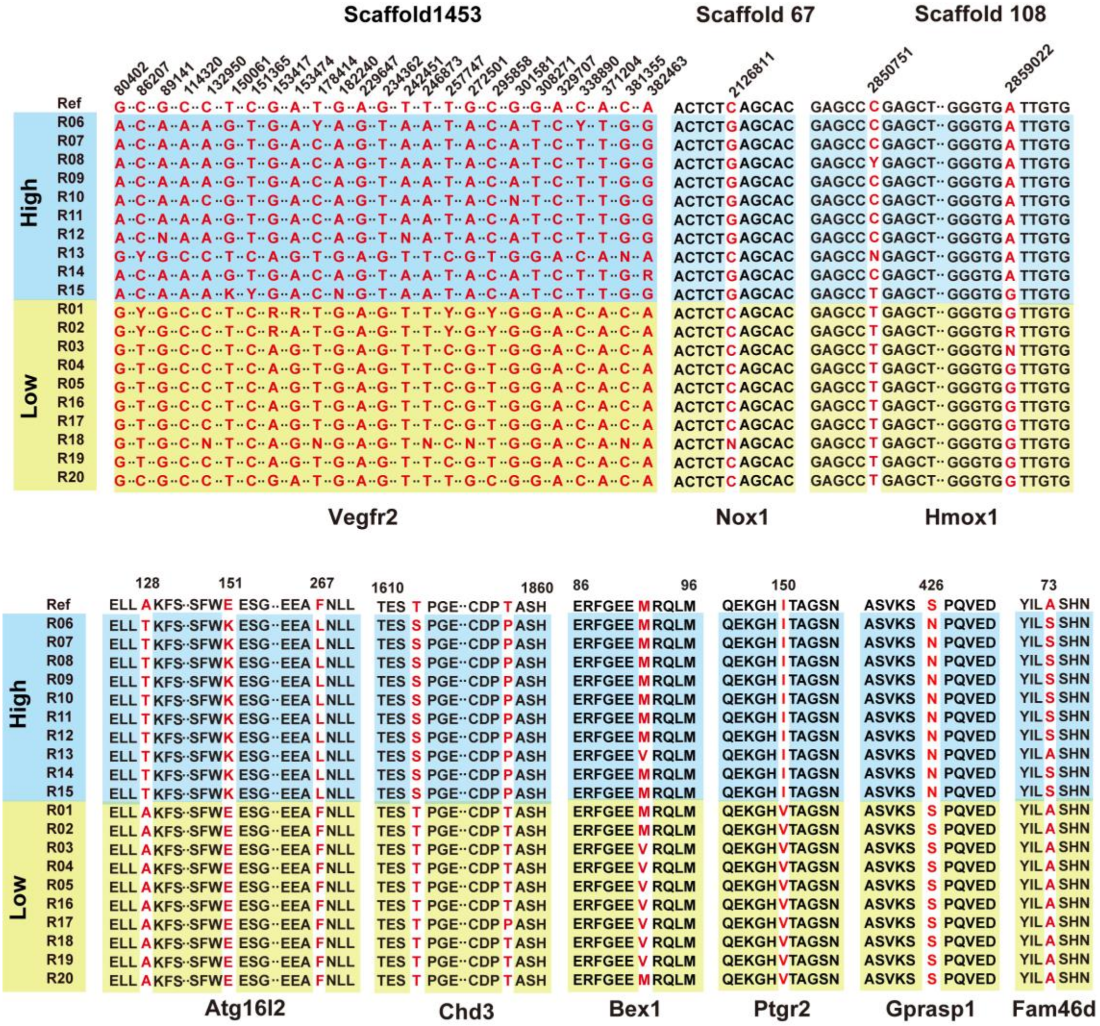
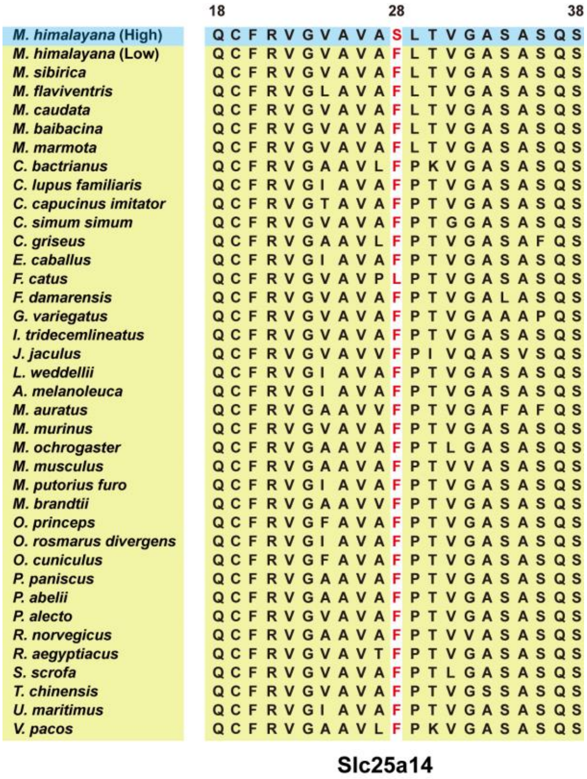


Figure S16. SNPs of nine genes involved in the HIF-1 pathway, Related to Figure 4.

Vegfr2, Nox1, and Hmox1 directly participate in the HIF-1 mediated pathway. Slc25a14 is involved in the “response to hypoxia”. Atg16l2, Bex1, Ptgr2, and Gprasp1 are associated with “apoptosis”. Chd3 and Fam46d participate in “DNA repair”. Genotyping differences are shown between high (in blue) and low (in yellow) altitude groups. SNPs and amino acid variants are highlighted in red. SNP coordinates are based on the reference genome of the Himalayan marmot. Protein coordinates are referred to as those of humans.



	18	28	38
M. himalayana (High)	Q C F R V G V A V A	S	L T V G A S A S Q S
M. himalayana (Low)	Q C F R V G V A V A	F	L T V G A S A S Q S
M. sibirica	Q C F R V G V A V A	F	L T V G A S A S Q S
M. flaviventris	Q C F R V G L A V A	F	L T V G A S A S Q S
M. caudata	Q C F R V G V A V A	F	L T V G A S A S Q S
M. baibacina	Q C F R V G V A V A	F	L T V G A S A S Q S
M. marmota	Q C F R V G V A V A	F	L T V G A S A S Q S
C. bactrianus	Q C F R V G A A V L	F	P K V G A S A S Q S
C. lupus familiaris	Q C F R V G I A V A	F	P T V G A S A S Q S
C. capucinus imitator	Q C F R V G T A V A	F	P T V G A S A S Q S
C. simum simum	Q C F R V G V A V A	F	P T G G A S A S Q S
C. griseus	Q C F R V G A A V L	F	P T V G A S A F Q S
E. caballus	Q C F R V G I A V A	F	P T V G A S A S Q S
F. catus	Q C F R V G V A V P	L	P T V G A S A S Q S
F. damarensis	Q C F R V G V A V A	F	P T V G A L A S Q S
G. variegatus	Q C F R V G V A V A	F	P T V G A A A P Q S
I. tridecemlineatus	Q C F R V G V A V A	F	P T V G A S A S Q S
J. jaculus	Q C F R V G V A V V	F	P I V Q A S V S Q S
L. weddellii	Q C F R V G I A V A	F	P T V G A S A S Q S
A. melanoleuca	Q C F R V G I A V A	F	P T V G A S A S Q S
M. auratus	Q C F R V G A A V V	F	P T V G A F A F Q S
M. murinus	Q C F R V G V A V A	F	P T V G A S A S Q S
M. ochrogaster	Q C F R V G A A V A	F	P T L G A S A S Q S
M. musculus	Q C F R V G A A V A	F	P T V V A S A S Q S
M. putorius furo	Q C F R V G I A V A	F	P T V G A S A S Q S
M. brandtii	Q C F R V G A A V V	F	P T V G A S A S Q S
O. princeps	Q C F R V G F A V A	F	P T V G A S A S Q S
O. rosmarus divergens	Q C F R V G I A V A	F	P T V G A S A S Q S
O. cuniculus	Q C F R V G F A V A	F	P T V G A S A S Q S
P. paniscus	Q C F R V G A A V A	F	P T V G A S A S Q S
P. abelii	Q C F R V G A A V A	F	P T V G A S A S Q S
P. alecto	Q C F R V G V A V A	F	P T V G A S A S Q S
R. norvegicus	Q C F R V G A A V A	F	P T V V A S A S Q S
R. aegyptiacus	Q C F R V G V A V T	F	P T V G A S A S Q S
S. scrofa	Q C F R V G V A V A	F	P T L G A S A S Q S
T. chinensis	Q C F R V G V A V A	F	P T V G S S A S Q S
U. maritimus	Q C F R V G I A V A	F	P T V G A S A S Q S
V. pacos	Q C F R V G A A V L	F	P K V G A S A S Q S

Figure S17. Multiple protein sequence alignment of Slc25a14 orthologs among 37 vertebrates, Related to Figure 4.

The amino acid variant is highlighted in red. Amino acid coordinates are referred to as those of humans.

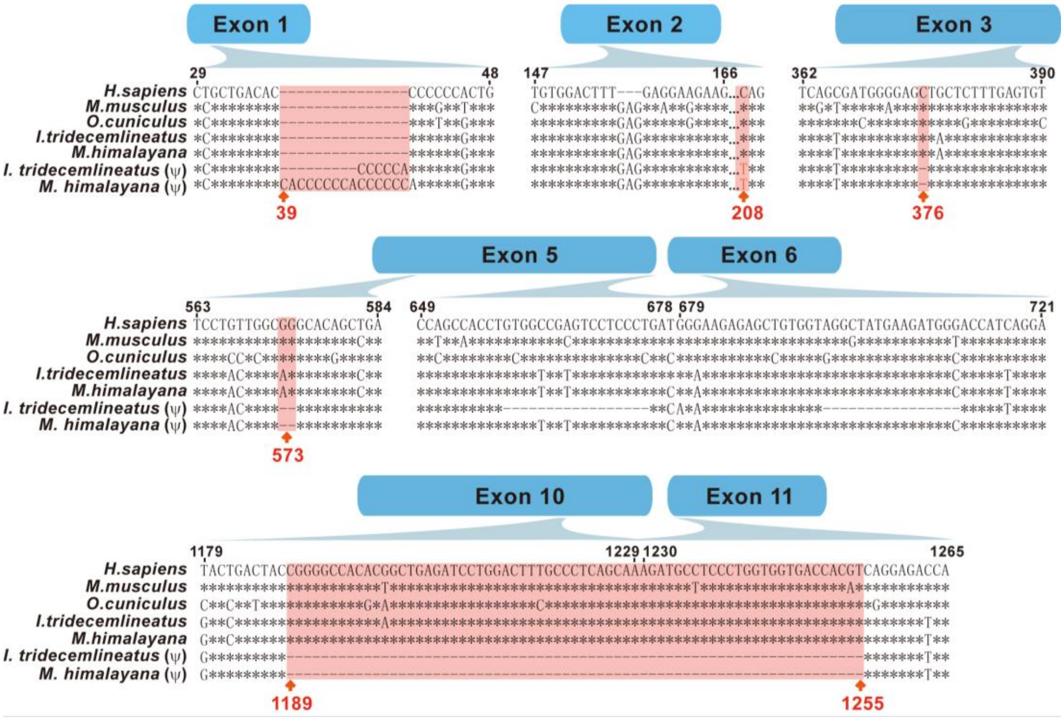


Figure S18. Multiple sequence alignment of ψAamp and Aamp, Related to Figure 5.

The positions of genetic variances of ψAamp in the Himalayan marmot are highlighted in light red. The blue box represents the corresponding exons at which ψAamp variants occurred. There is a 15-base pair insertion at position 39, one point-mutation at position 208 C>T causing a premature stop codon, two frameshift deletions at positions 376 and 573, and one 67-base pair deletion at position 1189. *H. sapiens*, human; *M. musculus*, mouse; *M. himalayana*, Himalayan marmot; *O. cuniculus*, rabbit; *I. tridecemlineatus*, squirrel.

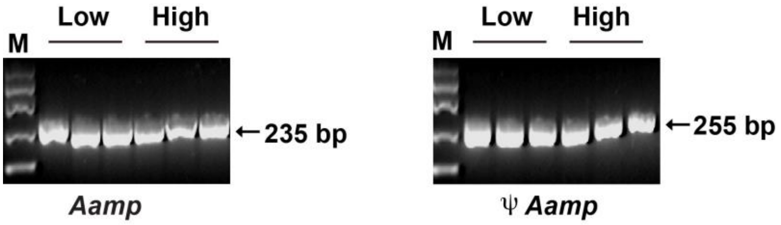


Figure S19. ψAamp and Aamp were detected in the liver, Related to Figure 5. Total RNA was extracted from liver tissues of high and low altitude Himalayan marmots. Semiquantitative PCR was used to detect the expression level of ψAamp and Aamp (n=3 for each group).

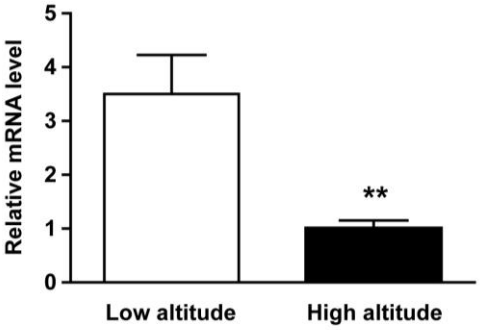


Figure S20. Aamp mRNA expression was decreased in the liver from high altitude Himalayan marmots, Related to Figure 5.

Total RNA was extracted from the livers of high- and low-altitude Himalayan marmots. Real time PCR was used to detect the expression level of Aamp (n=3 for each group).

Table S1. Summary of *de novo* sequencing data for the Himalayan marmot genome, Related to Figure 1.

Library insert size	Data (bp)	Coverage (X)	Q20 (%)	Q30 (%)
180 bp	44,494,699,378	18	98.03	92.59
180 bp	21,320,328,966	9	98.07	92.78
180 bp	62,530,484,561	25	94.24	90.28
500 bp	66,238,432,582	27	88.80	81.76
3 Kb	27,481,716,793	11	88.61	81.06
3 Kb	17,093,056,057	7	88.22	80.06
4 Kb	15,159,092,639	6	88.28	80.03
4 Kb	21,292,441,119	9	88.13	80.13
5 Kb	20,030,600,581	8	88.33	80.15
5 Kb	20,658,904,683	8	88.20	80.07
8 Kb	10,576,666,176	4	88.35	80.10
8 Kb	47,496,564,378	19	88.11	80.42
10 Kb	8,916,484,591	4	88.70	80.20
10 Kb	41,713,043,179	17	87.79	80.02
15 Kb	6,874,423,539	3	88.63	80.08
15 Kb	40,355,066,878	16	87.98	80.26
17 Kb	36,176,228,265	15	87.86	80.03
Total	508,408,234,365	206	89.78	82.35

Table S2. Summary of the Himalayan marmot genome assembly, Related to Figure 1.

Index	Contig	Scaffold*
N50 (Kb)	80.46	1,497.03
N90 (Kb)	21.1	381.46
Longest (Mb)	1.12	15.35
Total length (Gb)	2.36	2.47
Total number	58,723	4,566
Gap (Mb)	-	115.83 (4.69 %)

*Only scaffolds with more than 1 Kb were considered.

Table S3. Genome completeness evaluated by ESTs/unigenes, Related to Figure 1.

Range of length (bp)	Total number	Aligned number	Percent (%)
All	68,547	65,899	96.137
≥ 500	24,873	24,357	97.925
≥ 1,000	14,096	13,873	98.418

Note: only ESTs/unigenes with coverage ≥ 50% were calculated.

Table S4. Genome completeness assessed by CEG, Related to Figure 1.

Species	Number[a] (458 of CEGs)	Percentage[a]	Number[b] (248 of CEGs)	Percentage[b]
Himalayan marmot	456	99.6	247	99.6

Note：a. Number and percentage of 458 core eukaryotic genes (CEGs) included in the genome assembly. b. Number and percentage of 248 highly conserved CEGs included in the genome assembly.

Table S5. Genome completeness assessed by BUSCO, Related to Figure 1.

Categories	Number	Percent (%)
Complete BUSCOs	864	88.34
Complete and single-copy BUSCOs	768	78.53
Complete and duplicated BUSCOs	96	9.82
Fragmented BUSCOs	32	3.27
Missing BUSCOs	82	8.38
Total BUSCO groups found	896	91.62

Note: BUSCO, benchmarking universal single-copy orthologs.

Table S6. Summary of repeat contents of Himalayan marmot genome assembly, Related to Figure 1.

Type	Number	Length (bp)	Percent (%)
ClassI/DIRS	6,931	1,984,155	0.080
ClassI/LINE	2,153,488	495,713,192	20.07
ClassI/LTR	1,004,875	227,826,063	9.224
ClassI/LTR/Copia	795	63,769	0.003
ClassI/LTR/Gypsy	17,158	1,545,089	0.063
ClassI/PLE|LARD	498,083	78,401,712	3.174
ClassI/SINE	1,009,438	151,232,482	6.123
ClassI/SINE|TRIM	34	29,890	0.001
ClassI/TRIM	6,875	8,876,102	0.359
ClassI/Unknown	2,773	402,795	0.016
ClassII/Crypton	1,230	85,847	0.003
ClassII/Helitron	10,272	1,016,894	0.041
ClassII/MITE	3,497	528,481	0.021
ClassII/Maverick	8,722	857,800	0.035
ClassII/TIR	582,687	96,644,475	3.913
ClassII/Unknown	67,095	7,042,177	0.285
PotentialHostGene	5,332	1,705,472	0.069
SSR	139,767	43,098,496	1.745
Unknown	162,107	32,039,128	1.297
Total	5,681,159	1,149,094,019	46.52

Table S7. Summary of predicted non-coding RNA in Himalayan marmot genome, Related to Figure 1.

Class	Number	RNA family	Length (bp)
tRNA	467	-	38,272
rRNA	97	4	21,169
miRNA	291	178	32,232
snRNA	3,679	247	492,554
sRNA	107	18	17,970
lncRNA	264	176	37,949
Total	4,905	445	640,146

Note: tRNAs: transfer RNAs, rRNA: ribosome RNA, miRNAs: microRNAs, sRNA: small RNA, snRNAs: small nuclear and small nucleolar RNAs, lncRNA: long non-coding RNA.

Table S9. Summary of pooled transcriptome data assisted for genome annotation, Related to Figure 1.

Insert size (bp)	Read number	Read length (bp)	Size (bp)	GC (%)	N (%)	Q20 (%)	Q30 (%)
300	27,389,991	101	5,532,274,202	52.13	0.02	94.56	89.27

Note: A pooled transcriptome data of kidney, pancreas, adrenal gland, liver, and brain was used for genome annotation.

Table S10. Summary of predicted gene models, Related to Figure 1.

Strategy	Software	Gene number
De novo	August	33,718
Homology-based	Genewise + GeMoMa	22,791
	Tophat + Cufflinks	20,389
Transcriptome-based	Transdecoder	10,442
	PASA	21,786
Final gene set	EVM + manual	21,609

Table S11. Summary of gene models annotated by different databases, Related to Figure 1.

Category	Database	Gene number	Percent (%)
Annotated	EggNOG	13,978	64.7
	GO	17,391	80.5
	KEGG	12,892	59.7
	KOG	13,154	60.9
	Pfam	20,275	93.8
	Swissprot	21,423	99.2
	TrEMBL	18,213	84.3
	NR	19,646	90.9
	NT	20,212	93.6
	All	21,468	99.4
Unannotated		141	0.62
Total		21,609	100

Transparent Methods

Ethics approval

All experimental procedures and sample collection methods in this study involving Himalayan marmots were preapproved by the Animal Care and Use Committee of the Qinghai Association for Laboratory Animal Science and conducted according to the Wildlife Conservation Act (amendment on July 2nd, 2016, China) and Experimental Animal Management Regulations (amendment on March 1st, 2017, China).

Genome sequencing and assembly

A two-year-old male Himalayan marmot from Xining, Qinghai province, China, was used for *de novo* genome sequencing. DNA was extracted from the liver and blood using the Blood & Cell Culture DNA Mini Kit (Qiagen). DNA concentrations and quality were measured using a NanoDrop 2000 (Thermo) and a Qbit Fluorometer (Thermo Fisher), respectively. Library preparation and quality assessments were conducted according to the standard protocol of the HiSeq 4000 platform (Illumina, USA). Four short fragment paired-end libraries (mean insert sizes: 180 bp and 500 bp) and thirteen mate-end libraries (mean insert sizes: 3 Kb, 4 Kb, 5 Kb, 8 Kb, 10 Kb, 15 Kb, and 17 Kb) were constructed using the Illumina standard pipeline (Table S1). All libraries were sequenced on the Illumina HiSeq 4000 platform (Illumina, USA). Raw data were then filtered by the following criteria: (1) Filter paired-reads with an ambiguous nucleotide ratio greater than 5%; (2) Filter low-quality reads with a mean PHRED score < 20% (referring to sequencing error rates < 1%); (3) Filter reads with adapter contamination; (4) Filter duplication reads for mate-end libraries. A total of 508.41 Gb clean data was retained (Table S1) and then used for *de novo* assembly by ALLPATHS-LG(Maccallum et al., 2009) with default parameters. We then employed SSPACE(Boetzer et al., 2011) to connect contigs into scaffolds based on mate-end reads from long-insert jumping libraries (insert sizes ≥ 1 Kb). GapCloser (V1.12 for SOAP *de novo*)(Luo et al., 2012) was applied to fill gaps and improve the continuity of the scaffolds using pair-end reads (insert sizes < 1 Kb). The final assembled genome size was ~2.47 Gb. The scaffold N50 was 1.50 Mb with the longest scaffold length of 15.35 Mb, and the contig N50 was 80.46 Kb with the longest contig length of 1.12 Mb (Table S2). The assembled genome contained ~115.83 Mb gaps, which accounts for 4.69% of the assembly (Table S2). Furthermore, only scaffolds with lengths longer than 1,000 bp were used in further analyses.

Genome size estimation based on k-mer distribution

Genome sizes were estimated using JELLYFISH(Marcais and Kingsford, 2011) with an optimal k-mer size (http://koke.asrc.kanazawa-u.ac.jp/HOWTO/kmer-genomesize.html). The k-mer refers to an artificial division of k nucleotide sequences derived from an iterative cut of pair-end reads. Theoretically, the k-mer frequency follows a Poisson distribution. We selected k=21 for the genome size estimation in this study. Genome sizes were calculated from the following equation: Genome size = 21-mer number / 21-mer depth, where 21-mer number is the total counts of each unique 21-mer and 21-mer depth is the highest frequency that occurred. Consequently, the estimated genome size of the Himalayan marmot was ~2.33 Gb (Figure S1), which was close to the total scaffold length of the assembled genome. By taking the estimated genome size as a reference, total sequence data accounted for ~206-fold coverage (Table S1).

Genome assembly assessment

Transcriptomic data from a pooled RNA sample were used to quantify the integrity of the protein coding genome. Briefly, total RNA was extracted from the kidney, pancreas, adrenal gland, liver, and brain of the Himalayan marmot using Trizol Reagent (Life Technologies, California, USA) and was then mixed equally into a pooled RNA sample. Using the Illumina standard cDNA library construction and sequencing pipeline, we obtained 27,389,991 pair-end reads with an average read length of 300 bp (Table S9). After quality control, all reads were assembled into unigenes by Trinity(Grabherr et al., 2011) with default parameters. A total of 24,873 unigenes (length ≥ 500 bp) were obtained and then aligned back to the Himalayan marmot genome by BLAT(Kent, 2002). A total of 97.90% of the unigenes (length ≥ 500 bp) were supported by the reference genome (Table S3).

Core eukaryotic genes (CEGs), identified by the Core Eukaryotic Gene Mapping Approach (CEGMA v.2.3)(Parra et al., 2007), were mapped to the Himalayan marmot genome by BLAT(Kent, 2002) to assess the completeness of the protein coding region. We found 456 highly confident hits of CEGs, which accounted for 99.56% of all 458 CEGs. Moreover, the counterparts of 247 CEGs, accounting for 99.60% of 248 highly conserved CEGs, were found in the Himalayan marmot genome (Table S4). Additionally, the mammalian release of BUSCO v1.22 (mammalia_odb9)(Simao et al., 2015) was run on the genome of the Himalayan marmot, and it detected 864 complete gene models (including 768 single and 96 duplicated copies), making up 88.34% of the reference gene set, plus 32 fragmented gene models (Table S5).

Pair-end relationships and sequencing depth were examined to evaluate genome assembly accuracy. Five scaffolds were randomly selected for validation. The coherent pair-end relationship from 3 libraries (mean insert sizes: 4 Kb, 10 Kb, and 15 Kb) and high sequencing depth exhibited in Figure S2 suggested the high accuracy of the Himalayan marmot genome assembly.

The GC content and sequencing depth of the Himalayan marmot were calculated in each sliding window (window = 10 Kb, step = 1Kb) and then plotted to measure genome purity (Figure S3). We screened the genome sequence of 5 mammals (*M. himalayana*, *H. sapiens*, *M. musculus*, *O. cuniculus*, and *I. tridecemlineatus*) by a nonoverlapping sliding window (window = 10 Kb) for the GC content calculation and CpG island identification (Figure S4).

Repetitive sequence annotation

Homolog and *de novo* strategies were both applied to identify the repetitive sequence in the Himalayan marmot genome. Software, including RepeatScout(Price et al., 2005), LTR-FINDER(Xu and Wang, 2007), MITE(Han and Wessler, 2010), and PILER(Edgar and Myers, 2005), was used for *ab initio* prediction. The results obtained from the software were combined to form a new repetitive sequence database. This database was then merged with Repbase(Bao et al., 2015) and classified into different categories by the PASTEClassifier.py(Hoede et al., 2014) script included in REPET v2.5(Flutre et al., 2011). Repetitive sequences in the Himalayan marmot genome were identified by homolog searching with the final merged database through RepeatMasker(Chen, 2004). We identified ~1.149 Gb repetitive sequences, which accounted for 46.5% of the Himalayan marmot assembled genome (Table S6). Retrotransposon (Class I) comprised the majority of transposable elements, among which the long interspersed nuclear elements (LINE) and long terminal repeat (LTR) families were the two most abundant, accounting for 20.05% and 9.22% of the Himalayan marmot genome, respectively (Table S6).

Prediction of noncoding RNA (ncRNA)

The tRNA was predicted using tRNA-scan-SE (version 1.23) which utilized two embedded searching methods (tRNA-scan and EufindtRNA) and then analyzed by a highly selective tRNA covariance model(Lowe and Eddy, 1997). After filtering out the pseudo elements, the tRNAs were selected using a prediction score over 20 and repeat regions were masked out, resulting in a final set of 467 tRNA sequence. miRNAs were identified by homolog searching with one mismatch allowed using miRBase (Release 21)(Kozomara and Griffiths-Jones, 2014) as a reference. The secondary structures of putative sequences were predicted by miRDeep2(Friedlander et al., 2012). miRNAs with hairpin structures were considered to be reliable. Regarding other types of ncRNA, Infernal(Nawrocki, 2014) was applied by comparing the secondary structure (*e*-value ≤ 0.01) based on the Rfam database (release 12.0)(Gardner et al., 2009). In total, 63,768 ncRNA belonging to 6 types, including transfer RNAs, ribosome RNA, microRNAs, small RNA, small nuclear and small nucleolar RNAs, and long noncoding RNA, were identified in the Himalayan marmot genome (Table S7).

Prediction of pseudogenes

Proteins from the human, mouse, rabbit, and Himalayan marmot genomes were aligned to the Himalayan marmot genome with tBLASTN(Camacho et al., 2009) for candidate homolog region identification. Pseudogenes were then identified via GeneWise(Birney and Durbin, 2000) with a frame shift and/or premature stop code occurrence in the coding region. Additionally, genes that had introns missing from the ortholog/parent functional genes were identified as processed pseudogenes. After redundant filtering and manual inspection, 1,479 confident pseudogenes were identified in the Himalayan marmot genome (Table S8).

Protein-coding gene prediction and functional annotation

De novo, homology-based and transcriptome-based strategies were applied to predict protein-coding genes in the Himalayan marmot genome. Proteins from three well-annotated mammalian genomes (human: GRCh38.p7; mouse: GRCm38.p4; rabbit: OryCun2.0) were used to perform the homolog-based prediction by GeneWise(Birney and Durbin, 2000) and GeMoMa(Keilwagen et al., 2016). Regarding *de novo* prediction, we used Augustus(Stanke et al., 2006) with parameters trained by unigenes, which were assembled from pooled transcriptome data. As a third approach, unigenes, assembled from pooled transcriptome data, were aligned to the genome assembly using BLAT(Kent, 2002) (identity >= 0.95, coverage >= 0.90) and then filtered using PASA(Haas et al., 2003). We also mapped pooled transcriptome data to the reference genome using TopHat(Trapnell et al., 2012) and assembled transcripts with Cufflinks(Trapnell et al., 2012). Transdecoder(Grabherr et al., 2011) was then applied to identify the gene structure of new gene models and transcripts derived from Cufflinks. By giving the weights of the three methods, all predicted gene structures were integrated into consensus gene structures using EVidenceModeler (EVM)(Haas et al., 2008). To obtain reliable protein-coding gene models, the gene set was then filtered using the following steps: 1) Filter CDS lengths shorter than 300 bp; 2) Filter CDS whose lengths are not triple; and 3) Filter gene models with a premature stop codon. A total of 21,609 protein-coding genes were generated for the Himalayan marmot genome (Table S10).

Gene functions were assigned according to the best match of the alignments against various protein databases using BLASTP(Camacho et al., 2009) (*E*-value = 1e^{-5}), including the nonredundant protein (Nr) database, Swiss-Prot database, and TrEMBL. Furthermore, unigenes were searched against the NCBI nonredundant nucleotide sequence (Nt) database using BLASTN(Camacho et al., 2009) by a cut-off of *E*-value = 1e^{-5}. InterProScan (v4.3)(Quevillon et al., 2005) was used to collect domain information and GO term(Dimmer et al., 2012) annotation. To predict the most likely function of genes, all genes were aligned (*E*-value = 1e^{-5}) with KEGG proteins. Gene sequences were also aligned to the Clusters of Orthologous Group (KOG) database to predict and classify functions. Meanwhile, KAAS(Moriya et al., 2007) (KEGG Automatic Annotation Server) was used for the KEGG pathway annotation(Kanehisa and Goto, 2000). Functional annotation was further improved by a database of orthologous groups and functional annotation (eggNOG)(Huerta-Cepas et al., 2016). In total, 21,468 protein-coding genes were assigned by specific functions, accounting for 99.38% of the whole protein-coding gene set (Table S11).

Assessment of genome annotation quality

To assess the accuracy of protein-coding gene models, we compared the protein-coding gene structural characteristics of the Himalayan marmot genome with 3 well-annotated genomes (human: GRCh38.p7; mouse: GRCm38.p4; rabbit: OryCun2.0). As shown in Figure S5 and Table S12, the distributions of all parameters, including the total gene length, whole CDS length, single exon length, whole intron length, single intron length, and exon number per gene, were very similar between the Himalayan marmot and other species, indicating the high quality of the protein-coding gene set of the Himalayan marmot genome.

Orthologous and paralogous gene identification

All of the reference genome sequences and annotation sets were downloaded from the Ensemble database (release 56). The protein-coding gene models were used for the homolog prediction of Himalayan marmot. In consideration of alternative splicing variants, we selected the longest transcripts to represent the coding sequence. The treeFam(Li et al., 2006) methodology was reproduced to generate the gene family as a group of genes that descended from a single ancestral gene either by a speciation or duplication event(Vilella et al., 2009). We performed an all-to-all BLASTP(Camacho et al., 2009) comparison with an e-value cut-off of $1e^{-5}$. To weight the similarity of each gene pair, we assigned the H-score that ranged between 0 and 100, which was calculated by the equation: H-score = score (gene1 vs gene2) / max ((score (gene1 vs gene1), score (gene2 vs gene2)), where the score is the BLASTP bit score of each gene pair. We then built a hierarchy graph by hcluster_sg(Ruan et al., 2008), requiring the minimum edge (H-score) to be greater than 5 and the minimum edge density to be larger than 0.34 to form a cluster. The clustering of a gene family immediately stopped once there were more than one outgroup genes.

Gene families were subjected to multiple sequence alignments by the combined use of MUSCLE(Edgar, 2004a, Edgar, 2004b) and Mafft software(Katoh and Standley, 2013). Protein alignments were back-translated to nucleotides and unconfident regions were removed. We constructed phylogenetic trees using TreeBest (http://treesoft.sourceforge.net/treebest.shtml), which has a build-in algorithm to build the best tree that reconciled with the species tree and rooted with the minimized number of duplications and losses. Using phylogeny, a pairwise relationship (orthologues and within-species paralogues) could be inferred and one-to-one single copy orthologs detected using a customized Perl script.

Phylogenetic tree construction and divergence time estimation

In the phylogenetic analysis, single copy orthologs were identified following the procedure described in above section. A total of 4,573 single-copy orthologs, shared by 15 species (the mouse, rat, hamster, Himalayan marmot, squirrel, pika, rabbit, human, chimpanzee, goat, sheep, cattle, pig, donkey, and horse) were obtained. The protein sequences of single-copy orthologs were aligned by MUSCLE(Edgar, 2004a, Edgar, 2004b) and then concatenated into a super protein. We then constructed the phylogenetic tree using the maximum likelihood (ML) algorithm with the JTT amino acid substitution model implemented in phyML software(Guindon et al., 2010). The divergence time was estimated using the MCMCtree program in PAML (Phylogenetic Analysis of ML) package(Yang, 2007). Three calibration points (chimpanzee vs humans: 6.1~8.2 MYA, rat vs mouse: 12.6~22.2 MYA, root: 83.0~105.7 MYA) derived from the TimeTree database (http://www.timetree.org/)(Hedges et al., 2015) were applied to constrain the divergence time of the nodes.

Comparative genomic analysis

To define gene families that descended from a single gene in the last common ancestor, we downloaded the protein-coding genes of human, mouse, rabbit, pika, ground squirrel, yak and Tibetan antelope from the Ensemble database (release 56). The longest transcript was used to represent the gene. Treefam methodology was used to define gene families(Li et al., 2006). We then applied the likelihood model implemented in the software package Café(ref) to identify the expanded and contracted gene family along each branch of the phylogenetic tree(Yang and Nielsen, 2000). The topology and branch lengths of the phylogenetic tree were considered to infer the significance of change in gene family size (Figure 1B). The levels of significance for expansion and contraction were set at 0.05. The single copy orthologs of these species were also identified to detect the genes under positive selection in the Himalayan marmot genome. Then, we used the optimized branch-site model to compute the P values of likelihood ratio test assuming that the null model was a 50:50 mixture of a point mass at zero and the chi-squared distribution with 1 degree of freedom(Zhang et al., 2005). Meanwhile, we compared the gene families of human, mouse, rabbit and Himalayan marmot to determine the specific gene families of Himalayan marmot (Figure S6). The GO enrichment analysis was then applied to define the significantly enriched biological process of specific or expanded gene families and positively selected genes (Tables S13-S14, S16).

Samples for whole genome resequencing

Four *Marmota* species, including the long-tailed marmot (*M. caudate*), Mongolia marmot (*M. sibirica*), gray marmot (*M. baibacina*), and yellow-bellied marmot (*M. flaviventris*), were subjected to whole genome resequencing in order to reveal the genomic evolution of *Marmota*. The genome sequence and gene set (CDSs, proteins) of the Alpine marmot (*M. marmota*) were downloaded from NCBI with GenBank assembly accession: GCA_001458135.1. Detailed information is shown in Table S17. To uncover the plateau adaptation mechanisms of the Himalayan marmot, 20 individuals, including 10 (8 males and 2 females) from low altitude areas (~1,900 m, Huzhu, Qinghai province, China) and 10 (5 males and 5 females) from high altitude areas (~4,500 m, Yushu, Qinghai province, China), were caught in the field for whole genome resequencing (Table S19). To avoid close genetic relationships, 20 Himalayan marmots were collected from different villages. Peripheral blood samples were collected to measure routine blood indexes, including the hemoglobin level of the Himalayan marmots. Student’s t test (*t*-test) was applied to detect the significance between these two populations (Table S20).

DNA extraction and sequencing

Genomic DNA was extracted from the liver of each Himalayan marmot using the Blood & Cell Culture DNA Mini Kit (Qiagen). After quality control for DNA, we constructed 24 pair-end libraries with insert sizes of 500 bp following the Illumina standard pipeline. All libraries were sequenced on the Illumina 4000 platform and data quality control was implemented. An approximately 10-fold sequence depth was obtained for each sample. Detailed information is shown in Table S17 and Table S19.

SNP calling and filtering

Clean reads were mapped to the Himalayan marmot reference genome using BWA(Li and Durbin, 2009) and sorted by samtools(Li et al., 2009). We then marked duplicates with Picard tools (v1.94) (http://broadinstitute.github.io/picard/). SNPs and indels for each Himalayan marmot were called using the Genome Analysis Toolkit (GATK 3.0)(McKenna et al., 2010). In SNP calling, variants were retained unless they met any of the following criteria: 1) an overall quality (QUAL) score of < 30; 2) a mapping quality (MQ) score of < 60; 3) a phred-scaled p-value (FS) > 60; 4) a variant quality by depth (QD) score < 2; 5) genotype quality (GQ) score < 5. In total, 24,192,293 SNPs for five marmots were retained. In 20 Himalayan marmots, more stringent criteria were applied to filter SNPs with minor allele frequency < 0.05 and min integrity < 0.8, and 8,374,374 SNPs were then reserved for a plateau adaptation analysis (Table S17 and Table S19). We annotated SNPs using the SnpEff program(Cingolani et al., 2012) (Table S17 and Table S19).

Phylogenetic analysis of *Marmota* lineages

We selected the ground squirrel (accession number: GCA_000236235.1) as an outgroup to construct the phylogenetic tree of six marmots: Himalayan marmot, gray marmot, yellow-bellied marmot, long-tailed marmot, Mongolian marmot, and Alpine marmot. Single copy orthologs were identified among the Himalayan marmot, Alpine marmot, and ground squirrel following the method described above. Overall, 10,397 single copy orthologs were retained. We then reconstructed the corresponding CDS and protein sequences of all single copy orthologs for four other marmots based on the SNP sets from resequencing data. After multiple sequence alignment by MUSCLE (v3.349)(Edgar, 2004a), we removed gaps and bad alignment regions by a sliding window (protein: window=7 aa, step=1 aa; CDS: window=21 bp, step=3 bp; mismatch ratio ≥ 50%). The retained sequences were then concatenated into one super gene for each species. We subsequently constructed phylogenetic trees using the ML algorithm with the JTT amino acid substitution model implemented in phyML software(Guindon et al., 2010). Phylogenetic trees based on either CDSs or proteins exhibited the same topology, both supporting the Alpine marmot initially splitting from the 5 other marmots. The divergence time (7.7 ~ 11.5 MYA) of the ground squirrel and Himalayan marmot (Figure 1A) was adopted to constrain that of *Marmota* lineages. The diverged time was calculated using the MCMCtree program in the PAML package(Yang, 2007)(Figure 2B). Meanwhile, six genes located in the Y-chromosome were identified in Himalayan marmot and Alpine marmot taking mouse as reference. The corresponding CDS sequences of the other 5 marmots were constructed based on the SNPs. A phylogenetic tree was subsequently generated by the above method and showed the same topological relation as the one based on the whole genome sequence (Figure S8).

Heterozygosity estimation of *Marmota* lineages

Heterozygosity is defined as the rate of heterozygous SNPs along the genome. We calculated the heterozygosity rate in a nonoverlapping sliding window (window = 100

Kb) for all five marmots. The genome-wide distribution of heterozygosity was shown in a violin plot using ggplot2 of the R package (http://ggplot2.tidyverse.org/reference/scale_brewer.html) (Figure 2C).

Inference of the demographic history

We inferred the demographic histories of 5 marmot lineages and 3 Himalayan species (golden snub-nosed monkey, yak and giant panda) using the pairwise sequentially Markovian coalescent (PSMC) model(Tong et al., 2009) to diploid genome sequences (Figures S7 and S9). To exclude the influence of the different evolution ratios of sexual chromosomes/scaffolds in demographic inferring, we aligned the genomes of the Himalayan marmot, golden snub-nosed monkey (GenBank: GCA_000769185.1), yak (GenBank: GCA_000298355.1), and giant panda (GenBank: GCF_000004335.2) to the sex chromosomes of the rabbit (GenBank: GCA_000003625.1), rhesus monkey (GenBank: GCA_000772875.3), cow (GenBank: GCA_000003055.5), and dog (GenBank: GCA_000002285.2) using Mummer 3.0(Kurtz et al., 2004). Scaffolds/chromosomes with more than half of the length aligned to the sexual chromosomes were removed. Raw data for the golden snub-nosed monkey (SRA: SRP033389), yak (SRA: SRP059061) and giant panda (SRA: SRP000962) were downloaded from the SRA database of NCBI. The diploid sequence for PSMC input was generated by BWA(Li and Durbin, 2009) and SAMtools(Lagnel et al., 2009). To improve mapping quality, the parameter “-C” was set to be 50 and the regions of mapping coverage less than 1/3 or larger than 2-fold of the average coverage were filtered out. The autosome SNPs of five marmots were reserved to generate the diploid sequence for PSMC. The parameters for PSMC were set to the following: -p "4+25*2+4+6"; -t “15” -N “25”; and -r “5”. We estimated the mutation rate using the formula: $\mu = D \times g / 2 T$, where D is the observed sequence difference between two species, T is the estimated divergence time, and g is the estimated generation time. Sequence divergence between the Himalayan marmot and rabbit was estimated to be 14.0% using whole genome alignment with Mummer 3.0(Kurtz et al., 2004) and their divergence time (~82.5 MYA) was obtained from the TimeTree database (http://www.timetree.org)(Hedges et al., 2015). Thus, the mutation rate per generation per site (μ) was calculated to be 2.54×10^{-9}, which was very similar to that of mammals(Kumar and Subramanian, 2002). The same μ value and generation interval (g = 3) were applied for all five marmots. In the golden snub-nosed monkey(Pan et al., 2014), yak(Qiu et al., 2015), and giant panda(Zhao et al., 2013), the generation mutation rates (μ) were set as 1.36×10^{-8}, 5.84×10^{-9}, and 1.29×10^{-8}, while the generation intervals (g) were 10, 3, and 12 years, respectively. We applied a bootstrapping approach with repeated sampling 100 times to estimate the variance of the simulated results(Li and Durbin, 2011). The accumulation rate of the Chinese loess plateau (g/cm^3/1,000 years) was obtained from a previous study(Sun and An, 2005).

Genetic relationship and population structure

To infer the population structure of 20 Himalayan marmots, a principal component analysis (PCA) was performed using SMARTPCA in the EIGENSOFT 6.0 package with default parameters(Patterson et al., 2006) (Figure S13A). A population phylogenetic tree was constructed based on intragenic SNPs with full integrity using the HKY85 model in the ML algorithm(Guindon et al., 2010) (Figure S13B). The population structure was inferred using the program Admixture v1.22(Alexander et al., 2009). Genetic clusters (K) varying from 2 to 6 and cross−validation (CV) errors were performed to obtain the most likely K value. When K=2, there was a clear division between the high and low altitude groups with the lowest cross validation error (Figures S13C and S13D). Default settings were used to run Admixture v1.22.

Selective sweep regions for plateau adaptation

A 100-kb sliding window along with a step size of 10 kb was used to estimate population polymorphisms through the population fixation index (F_{st}) and nucleotide diversity (π) between high- and low-altitude populations. F_{st} and π values were calculated at each window using the BioPerl package(Stajich et al., 2002). Selective sweep regions with strong signals were defined as the combination of the top 5% of F_{st} (> 0.373) and the θ_π ratio (Low/High) (> 1.568) (Figure S14). The length of the selective sweep region was ~24.8 Mb (Table S21). We found 383 protein-coding genes and 27 pseudogenes located in these selective sweep regions (Table S22). We found 61 genes related to the HIF1 pathway, a key regulator of adaptive responses to reduced oxygen availability (Table S23).

Natural selection in selective sweep regions

CDS sequence sets were constructed based on the high-quality SNPs of 20 Himalayan marmots. To detect the selective pressure acting on the protein coding genes, we estimated the rate of nonsynonymous (Ka) and synonymous (Ks) (ω = Ka/Ks) substitutions site-by-site using the YN00 program with default parameters from the PAML 4.2b package(Yang, 2007). Each gene pair between high- and low-altitude individuals was estimated repeatedly. All Ka/Ks values were classified into two groups: selective sweep genes (SSGs) and nonselective sweep genes (NSSGs). A statistical analysis was performed using the Student's *t*-test. The results obtained are shown in the boxplot distribution (Figure S15).

Homology modeling of Slc25a14 proteins

The three-dimensional structures of Slc25a14 proteins from high and low altitude Himalayan marmots were predicted by I-TASSER(Roy et al., 2010). Homology modeling was based on its homolog, the apical sodium-dependent bile acid transporter (ASBT) from *Yersinia frederiksenii* (PDB ID: 4N7X)(Zhou et al., 2014). Using Bio3D(Skjaerven et al., 2016), an interactive tool for comparative analyses of protein structures, the best models were selected based on structural similarities, and a fitting procedure was conducted. Visualization and distance calculations were both performed with PyMOL1.8(DeLano, 2002) (Figure 4C).

Phylogenetic tree and insertion time of ψAamp

The full CDS sequences of the Aamp gene were extracted from the gff file of humans (NM_001302545.1), squirrel (XM_005330460.2), rabbit (XM_008259128.2), and mouse (NM_001190444.1). The squirrel ψAamp sequence was downloaded from NCBI (NW_004936624.1). We constructed an ML tree by MEGA 7.0(Kumar et al., 2016) using the Kimura 2-parameter model with a bootstrap value of 1,000 and default parameters. The sequence distances of Aamp and ψAamp were estimated by a dismat program in the EMBOSS package (v6.6.0.0)(Rice et al., 2000). We selected the Kimura and Tamura substitution correction methods for distance calculations (Table S24). We corrected the mutation rate of ψAamp using the formula: $\mu_{(\psi Aamp)} = D / 2T$, where D is the observed sequence difference in ψAamp between two species that accumulated after the split of two species and T is the estimated divergence time. The divergence time between the marmot and squirrel was set as 9.8 MYA. The mutation ratio of ψAamp was subsequently calculated as 1.78×10^{-9} per site per year ($\mu_{(\psi Aamp)} = 3.48\% / (2 \times 9.8 \times 10^{6})$). The insertion time of ψAamp was calculated using the formula: $T_{(\psi Aamp)} = (D_{(\psi Aamp)} - D_{(Aamp)} / 2) / 2\mu_{(\psi Aamp)}$, where $D_{(\psi Aamp)}$ is the distance of ψAamp and Aamp in either the Himalayan marmot or squirrel, $D_{(Aamp)}$ is the sequence distance of Aamp between the Himalayan marmot and squirrel, and $\mu_{(\psi Aamp)}$ is the mutation ratio of ψAamp per site per year. The estimated insertion time was 22.64 MYA for the Himalayan marmot and 25.40 MYA for the squirrel (Figure 5C).

Prediction of miRNAs binding to ψAamp

The flanking sequence (~50 bp) around the SNPs of ψAamp was used for the miRNA binding prediction through the miRBase website (http://www.mirbase.org/search.shtml). Three miRNAs, including hsa-miR-6739-5p, mmu-miR-6935-3p and hsa-miR-5006-3p, were found to bind to the 3'UTR region of ψAamp with the parameters setting as score > 60 and *e*-value < $10e^{-4}$. The target region of these three miRNAs (17 bp) was aligned to the Himalayan marmot genome using BLASTN(Camacho et al., 2009). Twenty-one miRNAs were retained with the following standards: 1) match length > 11 bp; 2) distance between two matches < 200 bp; and 3) reverse complementary between two matches. These miRNAs were then subjected to the RNAfold web server for secondary structure prediction with the default parameters (http://unafold.rna.albany.edu/?q=mfold/RNA-Folding-Form). If the target region (17 bp) is located in the stem loop region and the free energy (ΔG) is lower than 20, miRNAs were considered reliable. Finally, two miRNAs, named mhi-miRNA-6739 and mhi-miRNA-6935, were retained (Figure 5E).

RT-PCR analysis

Total RNA was extracted from Himalayan marmot livers using RNAiso Plus (TaKaRa). Reverse transcription involved 1 μg of total RNA with the PrimeScript RT reagent kit and gDNA eraser (TaKaRa). In order to obtain pure PCR products, two sets of primers were used to amplify the ψAamp DNA fragment in two separate runs of PCR. In the amplification of the Aamp fragment, normal semi-quantitative PCR was performed using one primer set. The PCR reaction involved 0.8 μl (10 μM) forward and reverse primers and 10 μl 2× PCR Green Mix for a final volume of 20 μl. Primer sequences are shown below.

Primer name	Forward primer	Reverse primers
Aamp	CCGCTGACACCCCCCCGCTG	AAGGTGACCTCACCATCAT
ψAamp-first	ATGGAGTACGAATCAGAAAG	GGGATCCAGGCTCACACAAA
ψAamp-second	GGCCACCGCTGACACCAC	AAGGTGACCTCACCATCAT

Sample collection for hibernation

Himalayan marmots used for hibernation RNA-seq were housed in a clean facility (12-hr light/dark cycle) and fed standard chow and water *ad libitum*. Each of the five marmots were sacrificed in torpor or arousal stage by an injection of ketamine hydrochloride (2 ml for each animal), and then the livers and brains were collected for RNA-seq.

RNA extraction and sequencing

Total RNA was extracted from the liver and brain using Trizol (also known as TRI reagent). RNA concentrations were measured using NanoDrop 2000 (Thermo). RNA integrity was assessed using the RNA Nano 6000 Assay Kit of the Agilent Bioanalyzer 2100 system (Agilent Technologies, CA, USA). mRNA was purified from total RNA using poly-T oligo-attached magnetic beads. Due to the low quality of its RNA, one brain sample was removed. Nineteen RNA libraries with an average insert size of 300 bp were generated using the NEBNext UltraTM RNA Library Prep Kit following recommendations by the manufacturer, Illumina (NEB, USA). Library quality was assessed on the Agilent Bioanalyzer 2100 system. The clustering of index-coded samples was performed on a cBot Cluster Generation System using the TruSeq PE Cluster Kit v4-cBot-HS (Illumina) according to the manufacturer’s instructions. Library preparations were sequenced on the Illumina HiSeq 4000 platform (Illumina, USA).

Identification of differentially expressed genes

After filtering low-quality reads, the clean reads were aligned to the Himalayan marmot genome using TopHat(Trapnell et al., 2012). Gene expression was quantified and normalized by Cufflinks(Trapnell et al., 2012) in RPKM (reads per million per kilo bases) and genes with expression values greater than 0.5 RPKM were considered to be expressed(Cloonan et al., 2008). Two criteria were applied to identify differentially expressed genes: 1) > 2-fold change in expression; 2) false discovery rate (FDR) < 0.05.

Data Availability

The sequence data reported in this paper is in the BioProject: PRJNA407692.

Supplemental References

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19, 1655-1664.

Bao, W., Kojima, K. K., and Kohany, O. (2015). Repbase update, a database of repetitive elements in eukaryotic genomes. Mob DNA 6, 11.

Birney, E., and Durbin, R. (2000). Using GeneWise in the Drosophila annotation experiment. Genome Res 10, 547-548.

Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D., and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27, 578-579.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. (2009). BLAST plus : architecture and applications. Bmc Bioinformatics 10.

Chen, N. (2004). Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics Chapter 4, Unit 4 10.

Cingolani, P., Platts, A., Wang Le, L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., and Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6, 80-92.

Cloonan, N., Forrest, A. R., Kolle, G., Gardiner, B. B., Faulkner, G. J., Brown, M. K., Taylor, D. F., Steptoe, A. L., Wani, S., Bethel, G., et al. (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat Methods 5, 613-619.

Delano, W. L. (2002). Unraveling hot spots in binding interfaces: progress and challenges. Curr Opin Struct Biol 12, 14-20.

Dimmer, E. C., Huntley, R. P., Alam-Faruque, Y., Sawford, T., O'donovan, C., Martin, M. J., Bely, B., Browne, P., Mun Chan, W., Eberhardt, R., et al. (2012). The UniProt-GO annotation database in 2011. Nucleic Acids Res 40, D565-570.

Edgar, R. C. (2004a). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5, 113.

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

Edgar, R. C., and Myers, E. W. (2005). PILER: identification and classification of genomic repeats. Bioinformatics 21 Suppl 1, i152-158.

Flutre, T., Duprat, E., Feuillet, C., and Quesneville, H. (2011). Considering transposable element diversification in de novo annotation approaches. PLoS One 6, e16526.

Friedlander, M. R., Mackowiak, S. D., Li, N., Chen, W., and Rajewsky, N. (2012). miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Res 40, 37-52.

Gardner, P. P., Daub, J., Tate, J. G., Nawrocki, E. P., Kolbe, D. L., Lindgreen, S., Wilkinson, A. C., Finn, R. D., Griffiths-Jones, S., Eddy, S. R., et al. (2009). Rfam: updates to the RNA families database. Nucleic Acids Res 37, D136-140.

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. Nat Biotechnol 29, 644-652.

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59, 307-321.

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith, R. K., Jr., Hannick, L. I., Maiti, R., Ronning, C. M., Rusch, D. B., Town, C. D., et al. (2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res 31, 5654-5666.

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., White, O., Buell, C. R., and Wortman, J. R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol 9, R7.

Han, Y., and Wessler, S. R. (2010). MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Res 38, e199.

Hedges, S. B., Marin, J., Suleski, M., Paymer, M., and Kumar, S. (2015). Tree of life reveals clock-like speciation and diversification. Mol Biol Evol 32, 835-845.

Hoede, C., Arnoux, S., Moisset, M., Chaumier, T., Inizan, O., Jamilloux, V., and Quesneville, H. (2014). PASTEC: an automatic transposable element classification tool. PLoS One 9, e91929.

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., Rattei, T., Mende, D. R., Sunagawa, S., Kuhn, M., et al. (2016). eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res 44, D286-293.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28, 27-30.

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30, 772-780.

Keilwagen, J., Wenk, M., Erickson, J. L., Schattat, M. H., Grau, J., and Hartung, F. (2016). Using intron position conservation for homology-based gene prediction. Nucleic Acids Res 44, e89.

Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. Genome Res 12, 656-664.

Kozomara, A., and Griffiths-Jones, S. (2014). miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res 42, D68-73.

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol 33, 1870-1874.

Kumar, S., and Subramanian, S. (2002). Mutation rates in mammalian genomes. Proc Natl Acad Sci U S A 99, 803-808.

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. Genome Biol 5, R12.

Lagnel, J., Tsigenopoulos, C. S., and Iliopoulos, I. (2009). NOBLAST and JAMBLAST: New Options for BLAST and a Java Application Manager for BLAST results. Bioinformatics 25, 824-826.

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, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754-1760.

Li, H., and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. Nature 475, 493-496.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078-2079.

Lowe, T. M., and Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25, 955-964.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1, 18.

Maccallum, I., Przybylski, D., Gnerre, S., Burton, J., Shlyakhter, I., Gnirke, A., Malek, J., Mckernan, K., Ranade, S., Shea, T. P., et al. (2009). ALLPATHS 2: small genomes assembled accurately and with high continuity from short paired reads. Genome Biol 10, R103.

Marcais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27, 764-770.

Mckenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.

Genome Res 20, 1297-1303.

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., and Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res 35, W182-185.

Nawrocki, E. P. (2014). Annotating functional RNAs in genomes using Infernal. Methods Mol Biol 1097, 163-197.

Pan, Y., Wu, H., Liu, S., Zhou, X., Yin, H., Li, B., and Zhang, Y. (2014). Potential usefulness of baculovirus-mediated sodium-iodide symporter reporter gene as non-invasively gene therapy monitoring in liver cancer cells: an in vitro evaluation. Technol Cancer Res Treat 13, 139-148.

Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23, 1061-1067.

Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. PLoS Genet 2, e190.

Price, A. L., Jones, N. C., and Pevzner, P. A. (2005). De novo identification of repeat families in large genomes. Bioinformatics 21 Suppl 1, i351-358.

Qiu, Q., Wang, L., Wang, K., Yang, Y., Ma, T., Wang, Z., Zhang, X., Ni, Z., Hou, F., Long, R., et al. (2015). Yak whole-genome resequencing reveals domestication signatures and prehistoric population expansions. Nat Commun 6, 10283.

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., and Lopez, R. (2005). InterProScan: protein domains identifier. Nucleic Acids Res 33, W116-120.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16, 276-277.

Roy, A., Kucukural, A., and Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 5, 725-738.

Ruan, J., Li, H., Chen, Z., Coghlan, A., Coin, L. J., Guo, Y., Heriche, J. K., Hu, Y., Kristiansen, K., Li, R., et al. (2008). TreeFam: 2008 Update. Nucleic Acids Res 36, D735-740.

Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31, 3210-3212.

Skjaerven, L., Jariwala, S., Yao, X. Q., and Grant, B. J. (2016). Online interactive analysis of protein structure ensembles with Bio3D-web. Bioinformatics 32, 3510-3512.

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G., Korf, I., Lapp, H., et al. (2002). The Bioperl toolkit: Perl modules for the life sciences. Genome Res 12, 1611-1618.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res 34, W435-439.

Sun, Y. B., and An, Z. S. (2005). Late Pliocene-Pleistocene changes in mass accumulation rates of eolian deposits on the central Chinese Loess Plateau. J Geophys Res-Atmos 110.

Tong, J. C., Lim, S. J., Muh, H. C., Chew, F. T., and Tammi, M. T. (2009). Allergen Atlas: a comprehensive knowledge center and analysis resource for allergen information. Bioinformatics 25, 979-980.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7, 562-578.

Vilella, A. J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2009). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res 19, 327-335.

Xu, Z., and Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 35, W265-268.

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

Yang, Z., and Nielsen, R. (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol 17, 32-43.

Zhang, J. Z., Nielsen, R., and Yang, Z. H. (2005). Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol 22, 2472-2479.

Zhao, S., Zheng, P., Dong, S., Zhan, X., Wu, Q., Guo, X., Hu, Y., He, W., Zhang, S., Fan, W., et al. (2013). Whole-genome sequencing of giant pandas provides insights into demographic history and local adaptation. Nat Genet 45, 67-71.

Zhou, X., Levin, E. J., Pan, Y., Mccoy, J. G., Sharma, R., Kloss, B., Bruni, R., Quick, M., and Zhou, M. (2014). Structural basis of the alternating-access mechanism in a bile acid transporter. Nature 505, 569-573.