Title Page

Comparative genomics and metagenomics analyses of endangered Père David's deer (*Elaphurus davidianus*) provide insights into population recovery

Xuejing Zhang[1]*, Cao Deng[2]*, Jingjing Ding[3]*, Yi Ren[4]*, Xiang Zhao[2]*, Shishang Qin[5], Shilin, Zhu[2], Zhiwen Wang[2], Xiaoqiang Chai[5], Huasheng Huang[4], Yuhua Ding[6], Guoqing Lu[7], Lifeng Zhu[1]†

Author affiliations:

[1]Nanjing Normal University, College of life Sciences, Nanjing 210046, China.

[2]PubBio-Tech Services Corporation, Wuhan 430070, China.

[3]Jiangsu Academy of Forestry, Nanjing, China.

[4]Shanghai Majorbio Bio-pharm Biotechnology Co. Ltd., Shanghai, China.

[5]DNA Stories Bioinformatics Center, Chengdu, 610021, China.

[6]Jiangsu Dafeng Milu National Nature Reserve, Dafeng, China.

[7]University of Nebraska at Omaha, Omaha, USA.

*X.Z., C.D., J.D., Y.T. and X.Z. contributed equally to this work.

†To whom correspondence should be addressed. E-mail: zhulf@ioz.ac.cn

Keywords:

Père David's deer, comparative genomics, metagenomes, selective sweeping, population recovery

Short tile:

Père David's deer (Milu) genome and gut microbiomes

Abstract

The milu (Père David's deer, *Elaphurus davidianus*) has become a classic example of how highly endangered animal species can be rescued. However, the mechanisms that underpinned this population recovery remain largely unknown. As part of this study, we sequenced and analyzed whole genomes from multiple captive individuals. Following this analysis, we observed that the milu experienced a prolonged population decline over the last 200,000 years, which led to an elongated history of inbreeding. This protracted inbreeding history facilitated the purging of deleterious recessive alleles, thereby ameliorating associated threats to population viability. Because of this phenomenon, milu are now believed to be less susceptible to future inbreeding depression occurrences. SNP distribution patterns confirmed inbreeding history and also indicated sign of increased and increasing diversity in the recovered milu population. A selective sweep analysis identified two outlier genes (*CTSR2* and *GSG1*) that were related to male fertility. Furthermore, we observed strong signatures of selection pertaining to the host immune system, including six genes (*SERPINE1*, *PDIA3*, *CD302*, *IGLL1*, *VPREB3*, and *CD53 antigen*), which are likely to strengthen resistance to pathogens. We also identified several adaptive features including the over-representation of gene families encoding for olfactory receptor activity, a high selection pressure pertaining to DNA repair and host immunity, and tolerance to high-salt swamp diets. Moreover, glycan biosynthesis, lipid metabolism, and cofactor and vitamin metabolism were all significantly enriched in the gut microbiomes of milu. We speculate that these characteristics play an important role in milu energy metabolism, immunity, development, and health. In conclusion, our findings provide a unique insight into animal population recovery strategies.

Introduction

Milu were once widely distributed in the swamps of East Asia, and they were predominantly found in China (**Figure 1AB, Supplementary Fig. S1**). This species was first introduced to west in 1866 by Armand David (Père David)(Cao 2005), and subsequently became extinct in its native China in the early 20th century(Cao 2005). Fortunately, between 1894 and 1901, Herbrand Arthur Russell (the 11th Duke of Bedford), acquired the few remaining deer (18 individuals) from European zoos. These individuals were nurtured at Woburn Abbey in England(Cao 2005) (**Figure 1C**) and the current world population was derived from this herd(Cao 2005). In the mid-1980s, 77 individuals were reintroduced to captive facilities in China(Cao 2005; Jiang and Harris 2008), and populations were established in Beijing, Dafeng, Tianezhou and Yuanyang (**Figure 1C**). Since then, the populations have rapidly expanded, and the milu have managed to overcome the genetic bottleneck of inbreeding. The repopulation of milu is now deemed a classic example of how a highly endangered species can be rescued. However, the mechanisms that underpin this population recovery remain largely unknown.

Results and Discussion

We sequenced and analysed the milu genome and performed whole-genome re-sequencing for five another individuals. The assembled genome (2.58 GB; ~114-fold coverage) had a scaffold N50 value of 2.85 Mb (**Supplementary Table S2**). Assembly quality assessment was performed by aligning the transcripts from *Odocoileus virginianus* (white-tailed deer, WTD) and *Cervus nippon* (Chinese Sika deer, CSD) to the scaffolds of milu (>93.9% and >97.6% coverage, respectively) (**Supplementary Table S3**) and a core eukaryotic gene set (>92.0% conserved genes). We observed that repetitive sequences occupied 39.84% of the whole assembly (**Supplementary Table S4-S5**), and 22,126 protein-coding genes were predicted by combining *de novo* and evidence-based gene predictions (**Supplementary Table S6**).

Milu had been raised in enclosures for more than 1,200 years, with supplementation occurring through the introduction of wild individuals(Li et al. 2011). This resulted in a prolonged genetic bottleneck with low resultant genetic diversity. Results generated using the Pairwise Sequentially Markovian Coalescent (PSMC) model(Li and Durbin 2011) validated this hypothesis (**Figure 1D, Supplementary Fig. S33**). After the Last Glacial Maximum (LGM, ~20 thousand years ago/KYA)(Yokoyama et al. 2000), it is likely that milu suffered from the effects of climate change, over-hunting and/or habitat loss. Indeed, milu populations diminished, and there was a tendency towards continuous decreases. This is further evidenced by fossil records and associated literary records(Cao 2005).

Reduced population sizes increase the opportunity for inbreeding. The protracted existence of small

populations along with more recent declines resulted in high levels of milu inbreeding. When related individuals mate, the offspring carry long stretches of homozygous genome. Thus, the detection of runs of homozygosity (ROH) is a practical approach for estimating inbreeding at the individual level(Kim et al. 2013; Zhou et al. 2014) (**Supplementary Table S36**). When compared with 34 giant panda genomes(Zhao et al. 2013), 18 polar bear genomes(Liu et al. 2014) and eight Crested ibis(Li et al. 2014) genomes, we observed that the *Froh* (ROH length / Genome effective length) of milu ranged from 0.11 to 0.16. These values are much higher than those exhibited by the well-known panda (from 0.04 to 0.10) and polar bear (from 0.004 to 0.064), which are less prone to occurrences of inbreeding. However, the milu *Froh* values are lower than those exhibited by the previously critically-endangered crested ibis (from 0.19 to 0.32), which experienced a more recent and severe genetic bottleneck(Li et al. 2014) (**Figure 1E**). Length distribution of ROH also provides information about the timing of major inbreeding events. Long ROH are most likely derived from a recent ancestor; shorter ones, from a more distant ancestor(Curik et al. 2014). As revealed in **Figure 1F**, the milu has a medium average ROH length when compared with the crested ibis, the panda and the polar bear. The crested ibis contains an elongated ROH (longer than 1M), which is consistent with the fact that current crested ibis populations are derived from seven individuals approximately 40 years ago(Li et al. 2014). The milu harbors an increased average ROH length compared with the pandas and polar bears; however, this value is shorter than those observed for crested ibis. This would suggest that the time of major milu inbreeding event occurred prior to that of crested ibis but after those of panda and polar bear. These data confirm the existence of a prolonged reduced milu population.

Another major threat to small and endangered populations involves the loss of genetic diversity(Frankham 2005; Steiner et al. 2013). Small populations are susceptible to genetic drift and fixation, and these phenomena can be accelerated by inbreeding(Saccheri et al. 1998; Keller and Waller 2002; Steiner et al. 2013). We observed that genetic diversity was lower in the milu than in the panda, with a heterozygosity rate of 0.51 per kilobase pair in the milu, versus 1.32 per kilobase pair in the panda (**Supplementary Table S25**). Comparison with other endangered animals that experience, or have experienced, ongoing or recent population bottlenecks, indicated that this value was similar to that of mountain gorillas (Xue et al. 2015) (0.64×10^{-3}) but slightly higher than that of the crested ibis (0.36×10^{-3}, **Figure 2A**), Chinese alligator(Wan et al. 2013)(0.15×10^{-3}) and baiji(Zhou et al. 2013) (0.12×10^{-3}). In addition, patterns of SNP density distributions were explored by fitting a two-component mixture model to the observed SNP densities using the expectation-maximization algorithm(Hacquard et al. 2013) (**Figure 2B, Supplementary Table S30-S33**).

Half of the milu genome harbored only less than 5% of the called SNPs, and the mean heterozygosity of these low SNP density regions was 0.03 per kilobase, a value that was similar to that observed in crested ibis but much lower than those observed in panda and polar bear, reflecting more recent inbreeding history in milu and crested ibis. However, the mean heterozygosity in the other half of the milu genome was 1.26 per kilobase, which was similar to that observed in panda but higher than that observed in crested ibis, indicating a stronger sign of increased diversity in the recovered milu population than crested ibis population. Generally, the occurrence of heterozygosity in exons is reduced due to selective constraints(Li et al. 2014). However, the ratio of exon heterozygosity to genome heterozygosity in the milu and crested ibis is higher than that observed for the panda and polar bear (**Figure 2C, Supplementary Table S25)**. There are two possible explanations for this finding. First, it is possible that the milu and crested ibis experienced a slower rate of loss of genetic diversity in exons during inbreeding. Second, a rapid increase in the diversity of exons in recovered milu and crested ibis populations, following the occurrence of severe genetic bottlenecks, may have resulted in greater genetic diversity in these genetic regions. Inbreeding depression is a major force affecting the evolution and viability of small populations in captive breeding and restoration programs(Saccheri et al. 1998; Keller and Waller 2002; Steiner et al. 2013). Deleterious mutations tend to accumulate in associated populations due to reduced selective strength(Saccheri et al. 1998; Steiner et al. 2013). We observed that the milu exhibits a relatively low percentage of deleterious variants compared to other healthy or recovered populations (**Figure 2D**). This is consistent with a low effective population size (Ne) and the occurrence of inbreeding(Xue et al. 2015). In these populations, alleles occur more frequently in the homozygous state, and because deleterious variants are more likely to be pronounced, they are less likely to persist in the population (even if recessive)(Xue et al. 2015). Therefore, populations, such as the milu, that have experienced reduced population sizes for prolonged periods may be less susceptible to future inbreeding depressions because they have been purged of deleterious recessive alleles. Consequently, these populations are more likely to recover from future severe genetic bottlenecks.

Because of the prolonged history of captivity, reduced population size, and inbreeding associated with the milu, the study of adaptive evolution following exposure to these conditions is imperative in our efforts to prevent further future bottlenecks. We investigated adaptive evolution in the milu by analyzing the composition of several protein domains, and the expansion and contraction of a number of gene families[14,19,20]. We also investigated lineage-specific accelerated evolving GO categories(Sequencing and Consortium 2005; Bakewell et al. 2007; Qiu et al. 2012) and PSGs(Qiu et al. 2012; Zhou et al. 2013; Yim et al. 2014)

(**Supplementary Materials**). A functional analysis of the milu-specific expansion domains (**Supplementary Table S13**) showed that a large proportion of such domains is related to translation machinery. Notably, HSP90 genes in milu show a remarkable expansion in cytosolic members (*HSP90AA* and *HSP90AB*), especially the inducible HSP90AA1 and HSP90AA2 forms (**Supplementary Table S14, Supplementary Fig. S15**). The Hsp90 protein (PF00183) is important in stress response and has a capacity to buffer underlying genetic variation(Yeyati et al. 2007). Upon analysis of gene family numbers, we identified 835 and 4,584 gene families that expanded and contracted in the milu, respectively. In other mammals, it was observed that 77 gene families expanded ($p<0.01$, **Figure 3A**). The more pronounced expanded families were significantly over-represented (**Supplementary Table S11**) by genetic elements pertaining to 'olfactory receptor activity' ($P=3.29x10^{65}$), detection of chemical stimulus involved in sensory perception of smell ($P=2.00x10^{47}$), 'ATPase activity' ($P=6.26x10^{7}$), 'platelet dense granule membranes' ($P=5.65x10^{14}$), chloride channel activity ($P=7.92x10^{6}$), antigen processing and presentation of peptide antigen via MHC class I ($P=3.54x10^{3}$), cellular response to interferon-gamma ($P=1.35x10^{3}$), sperm mitochondrial sheath ($P=9.85x10^{3}$). These functional groups might play important roles in milu's behavior, development, immune and breeding. For example, much of the cellular response to interferon-gamma can be described in terms of a set of integrated molecular programs underlying well-defined physiological systems; and the induction of efficient antigen processing for MHC-mediated antigen presentation, which play clearly defined roles in pathogen resistance(Boehm et al. 1997). We also identified 26, 25, and 17 GO categories that demonstrated a significantly elevated pairwise number of non-synonymous substitution (*A*) values in the milu in the comparison of milu-cow-human, milu-TA-human and milu-baiji-human, respectively; while 14, 21, and 30 GO categories were elevated in cow, TA and baiji, respectively. In reference to the milu, the accelerated evolving GO categories were predominantly found to be involved in DNA repair, gene expression, protein modification, development, immunity, excretion, and responses to insulin stimuli (**Figure 3B, Supplementary Table S17-S18, Supplementary Fig. S17-S18**). Furthermore, 455 PSGs were identified using the likelihood ratio test implemented in PAML(Yang 2007) (**Supplementary Table S19**). These PSGs were enriched for genes involved in DNA repair, RNA metabolic processes, cellular protein modification processes, nitrogen compound metabolic processes, TLR 3 signaling pathways, regulation of development processes, and regulation of cytokine production (**Supplementary Table S20, Supplementary Fig. S19**).

In small captive populations, genetic adaptation to artificial environments can also occur, through processes including selective sweeps(Rubin et al. 2010; Rubin et al. 2012). We searched the genome for regions with

high degrees of fixation, and the distributions of observed *Hp* values and the Z transformations of *Hp*, *ZHp*, are shown in **Figure 3C**. In the genome-wide screen, 30 distinct gene loci showed a *ZHp* value lower than −6. Among the outliers derived following this analysis, we observed two genes that are related to male fertility, *CTSR2* (a.k.a. *CATSPER2*, cation channel sperm-associated protein 2) and *GSG1* (Germ cell-specific gene 1 protein). *CTSR2* complexes with other family members to form a calcium permeant ion channel, which plays a primary role in the regulation of sperm motility(Quill et al. 2003). *GSG1* colocalized with testis-specific poly(A) polymerase (*TRAP*) during spermiogenesis, and the interaction between *TPAP* and *GSG1* may be related to morphological alterations that occur during spermiogenesis (the transformation of round spermatids to elongating spermatids)(Choi et al. 2008). This may imply that potential selection of breeding stocks occurred in the milu population, thereby supporting the prolonged captive history of the latter. Interestingly, the gene family of sperm mitochondrial sheath ($P=9.85x10^{-3}$) was significant expanded in Milu genome. The mature sperm tail has several accessory structures, including a mitochondrial sheath, outer dense fibers and a fibrous sheath, and (Holstein 1976). Studies with gene knockout mice have proven that precisely regulated mitochondrial sheath formation is critical for sperm motility and fertility(Bouchard et al. 2000; Miki et al. 2004).

We also observed strong signatures of selection in relation to host immunity, including six genes (*SERPINE1*, *PDIA3*, *CD302*, *IGLL1*, *VPREB3*, and *CD53 antigen*), which may strengthen host resistance to pathogenic infection. Another interesting signature of positive selection was the *TAS2R* locus (**Figure 3C**). The *TAS2R* locus controls bitter taste sensitivity, including sensitivity to saccharin, quinine, and salicin(Deshpande et al. 2010). Moreover, we also found the significant gene family expansion on chloride channel activity in Milu genome, which mediates salt and liquid movement(Sheppard and Welsh 1999). By scanning milu-specific single amino acid polymorphisms (SAPs) in salt-sensitive ENaCs (epithelial sodium channels)(Chandrashekar et al. 2010), we identified 14 SAPs associated with *SCNN1A*, *SCNN1B*, *SCNN1G*, and *SCNN1D* *(***Supplementary Table S39, Supplementary Fig. S34-S37***)*. Eight SAPs were predicted to influence channel function, thereby affecting salt-sensation and sodium absorption (**Figure 3D**). Historically, milu were widely distributed in the eastern coastal regions of China(Cao 2005) (**Figure 1A**). Currently, the largest captive and wild release populations live in Dafeng Natural Reserve, in the eastern coastal shoal region of China (**Figure 1C**). The salinity of the main diet of these individuals is significantly higher than for inland populations (**Figure 3E, Supplementary Table S40-S41)**. Thus, the occurrence of polymorphisms in loci that are related to bitter and salt tasting sensations may explain the adaption of the milu to high-salt diets in swamp.

Symbiotic gut microbes play important roles in host nutrition, development, immunity, and health in animals(Ley et al. 2008). Metagenomic analysis of 10 milu gut microbial genomes and 39 mammalian microbial genomes (including whale, dolphin, carnivore, omnivore and herbivore genomes)(Muegge et al. 2011; Sanders et al. 2015) was performed using the MG-RAST online server(Meyer et al. 2008) (**Supplementary Table S42**). This analysis revealed functional enrichment of sodium transportation in milu gut microbes. Factors that were affected by this phenomenon included the Sodium transport system ATP-binding protein, Adenosinetriphosphatase, and Transcriptional regulatory protein NatR (**Figure 4A-C**). These occurrences may reflect an adaptation to a high salinity diet. Moreover, glycan biosynthesis, lipid metabolism, cofactor and vitamin metabolism (including folate biosynthesis, thiamine biosynthesis and vitamin B6 metabolism), and biosynthesis of other secondary metabolites (including penicillin and cephalosporins) were also significantly enriched in milu gut microbes (**Figure 4D-G**). It is possible that these reactions participate in host immunity, development, and health.

Acknowledgments This work was supported by grants from the National Natural Science Fund for outstanding young fund (31222009), National Natural Science Fund (31570489) and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

Author Contributions L.Z. conceived the study, L.Z. headed and Y.R managed the sequencing project, X.Z., and J.D prepared sequencing data, L.Z., C.D. and Z.W. coordinated the bioinformatics activities, L.Z., C.D., X.Z., S.Z., Z.W., S.Q. and X.C. designed experiments and analyzed the data, S.H., G. L., and Y.D. participated in project design, L.Z., C.D. and G. L. wrote and edited the manuscript with input from all other authors. All authors have read and have approved the manuscript.

Author Information The E. davidianus whole-genome sequences are deposited in GenBank under accession number JRFZ00000000.The 10 metagenomes of Milu gut microbes were submitted to MG-Rast, and the accession number were 4693474.3, 4693473.3, 4693472.3, 4693453.3, 4693450.3, 4693448.3, 4693446.3, 4693445.3, 4693207.3 and 4693196.3.Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to L.Z. (zhulf@ioz.ac.cn).

Competing financial interests The authors declare no competing financial interests.

Figure Legends

Figure 1. History of milu.

A, Palaeogeographic distribution history of wild milu in China. The data for milu fossils were adopted from Cao[1]. The color relates to the density of the fossils in specific provinces, and the density was calculated as the number of fossils per million square kilometers. **B**, Forage selection in coastal shoal habitat of milu in Dafeng Milu Natural Reserve, Jiangsu, China. **C**, Large-scale reintroduction programs since 1985. C, fawn; F, females; M, males. **D**, Demographic history of the milu. The history of the milu population and climate change spans from 3 KYA to 4 MYA. We used the default mutation rate of 1.5×10^{-8} for baiji (μ) and an estimation of 6 years per generation (g). The last glacial maximum (LGM) is highlighted in grey. Tsurf, atmospheric surface air temperature; RSL, relative sea level; 10 m.s.l.e., 10 m sea level equivalent. **E**, Box plot of *Froh* for milu, crested ibis, panda, and polar bear populations. *Fron* denotes the proportion of total ROH length. **F**. Box plot of length of ROH in each individual from milu, crested ibis, panda, and polar bear.

Figure 2. Genetic diversity of milu and other animals.

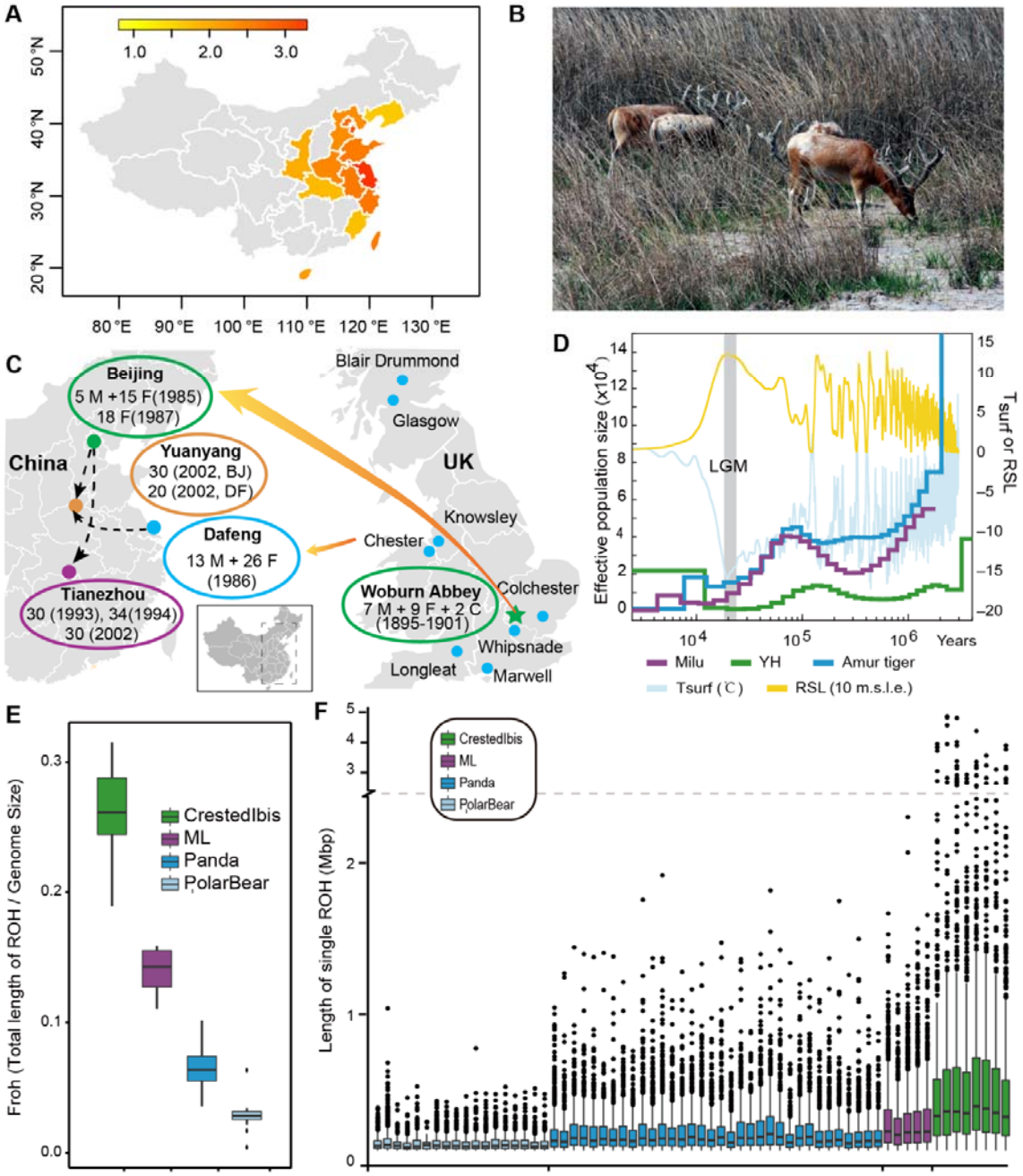
A. Box plot of heterozygosity from milu, crested ibis, panda, and polar bear individuals. Only heterozygous SNPs were included. CI, Crested Ibis; ML, Milu; PA: Panda; PB: Polar bear. **B**. Bias distribution of SNPs in animal genomes. Each circle denotes one species as (**A)**. L, low SNP density region; H, high SNP density region; kbp, kilobase; the proportion of total length of L and H regions in whole genome are green and purple; the proportion of SNP number in L and H region to total SNP number in both L and H regions are light blue and blue. **C**. Ratio of heterozygosity in each genomic element. The genomes were subdivided into three regions – exons, introns and other (regions that were neither exons nor introns). Then, heterozygosity in each type of genomic element was compared to heterozygosity of whole genome. **D**. Classification of missense variants. DE: deleterious; TO: Tolerated; and OT: Other.

Figure 3. Adaptive evolution in the milu genome.

A. Phylogenetic position of milu relative to other mammals. The branch lengths of the phylogenetic tree are scaled to demonstrate divergence time. Tree topology is supported by a posterior probability of 1.0 for all nodes. The blue bars on the nodes indicate the 95% credibility intervals of the estimated posterior distributions of the divergence times. The red circles indicate the fossil calibration times used for setting the upper and lower bounds of the estimates. The number of significantly expanded (green) and contracted (orange) gene families is designated on each branch. MRCA, most recent common ancestor. **B**. Lineage-specific accelerated evolving GO categories of biological process using the number of non-synonymous substitutions. **C**. Summary of selective sweep analysis. The negative end of the *ZHp* distribution presented along pseudo-chromosomes 1–29. The horizontal dashed lines indicate the threshold at $ZHp = -6$. Genes residing within 20 kb of a window with $ZHp \leq -6$ are indicated by their gene names. **D**. Red dot, milu-specific SAPs (single amino acid polymorphisms); red circle, damaging milu-specific SAPs predicted by PPH2. **E**. The salinity of forage plants in Dafeng Milu Natural Reserve.

Figure 4. The comparative metagenomic analysis of 10 milu gut microbial genomes and another 39 mammalian genomes (including genomes from whales, dolphins, carnivores, omnivores and herbivores). **A-C**, the genes coding for putative enzymes related to the sodium transport system, including Sodium transport system ATP-binding protein, Adenosinetriphosphatase, and Transcriptional regulatory protein, NatR. **D-G**, the genes coding for putative metabolism of cofactors and vitamins (folate biosynthesis, thiamine biosynthesis and vitamin B6 metabolism), and biosynthesis of other secondary metabolites (including penicillin and cephalosporin biosynthesis). CA, carnivores. WD, whales and dolphins. HE, herbivores. OC, omnivores. The number in brackets represents sample size.

292 Figure 1.

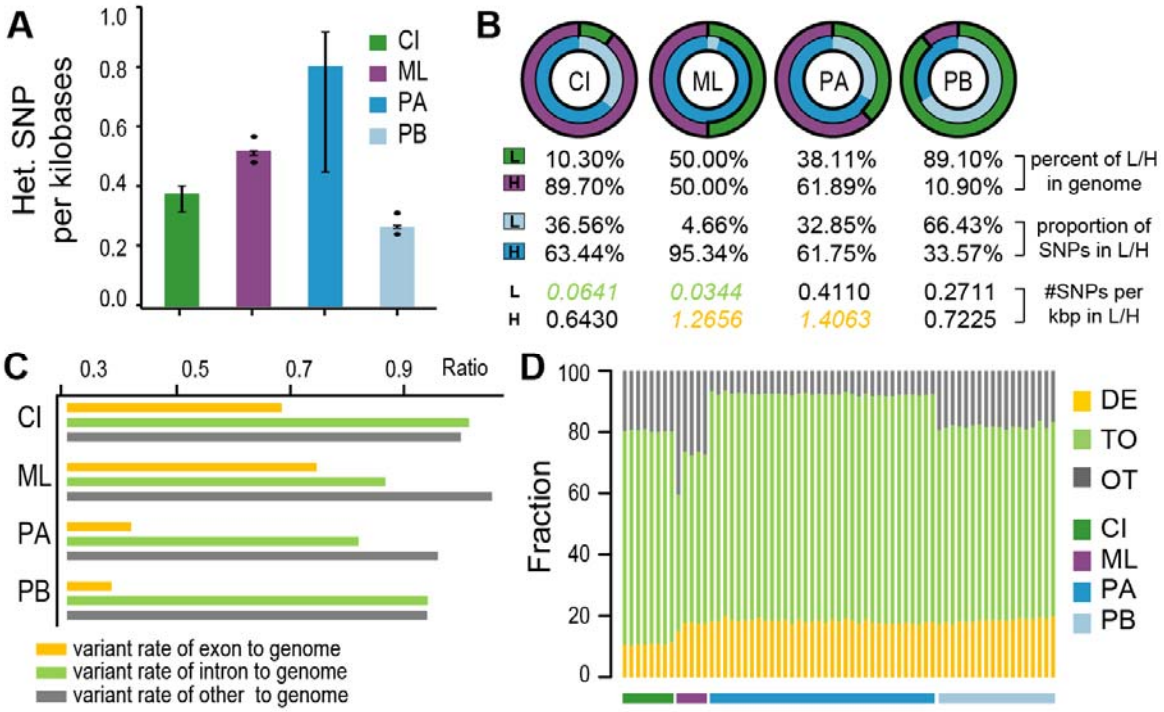


293

294

295 Figure 2.

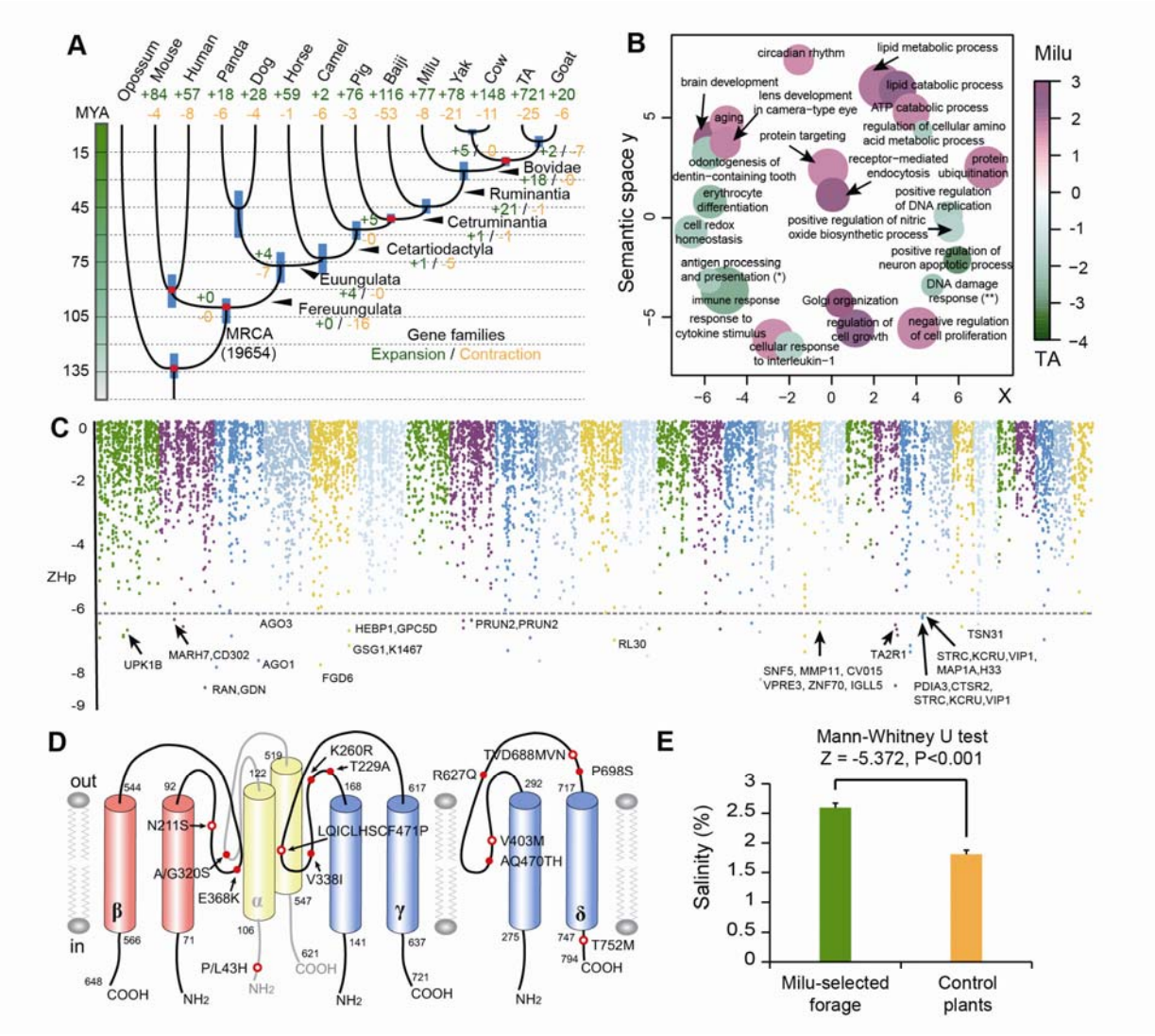
296



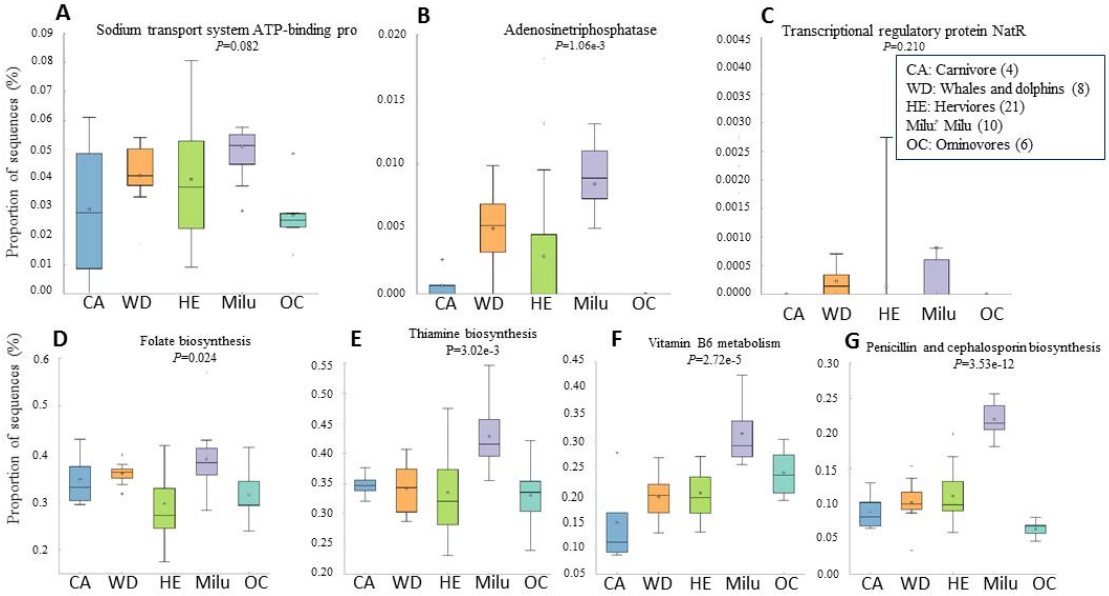
297

298

Figure 3.



302 Figure 4.



303

304

305

306

307

Online Methods

Genome sequencing and assembly

DNA from blood samples acquired from an adult female milu in Dafeng Milu Natural Reserve was used for *de novo* sequencing. Samples from an additional five animals were utilized for resequencing. Libraries with different insert sizes were constructed at Majorbio (Shanghai), and the insert sizes of the libraries were 180 bp, 500 bp, 800 bp, 3 kb, 5 kb, 8 kb, and 10 kb. The libraries were sequenced using a HiSeq2000 instrument. The other five resequencing samples were sequenced with read and insert lengths of 101 bp and 500 bp, respectively.

Whole-genome shotgun assembly of the milu was performed using the short oligonucleotide analysis package, SOAP*denovo*(Li et al. 2010). After filtering the reads, short-insert size library data were used to construct a *de Bruijn* graph without paired-end information. Contigs were constructed by merging the bubbles and resolving the small repeats. All qualified reads were realigned to contig sequences and paired-end relationships between the reads of allowed linkages between the contigs. We subsequently used the relationships, step by step, from the short-insert size-paired ends and the long-distance paired-ends to construct scaffolds. Gaps were then closed using the paired-end information to retrieve read pairs in which one end mapped to a unique contig and the other was located in the gap region. Assembly quality was assessed by aligning the assembled WTD(Malenfant et al. 2014) and CSD(Yao et al. 2012a; Yao et al. 2012b) transcripts with the milu scaffolds and by using a core eukaryotic gene mapping method(Parra et al. 2007).

Genome annotation

Transposable elements in the milu genome were identified by a combination of homology-based and *de novo* approaches**.** Tandem repeats were identified using Tandem Repeat Finder(Benson 1999). Interspersed repeats were characterized by homolog-based identification using RepeatMasker open-4.0.3(Smit et al. 1996) and the repeat database, Repbase[2]. Repeated proteins were identified using RepeatProteinMask and the transposable elements protein database. *De novo* identified interspersed repeats were annotated using RepeatModeler(Price et al. 2005), and LTR_FINDER(Xu and Wang 2007) was used to identify the LTRs; these results were used to generate the *de novo* repeat libraries, and then RepeatMasker was run once more against the *de novo* libraries. All repeats identified in this manner were included in the total count of interspersed repeats.

The milu protein-coding genes were annotated following the use of a combination of homolog gene prediction and *de novo* gene prediction tools. For homolog gene prediction, the protein sequences from cow, yak, goat, TA, and human were mapped to the genome using tBLASTn(Altschul et al. 1990), and

GeneWise(Birney et al. 2004) was used to predict the gene model based on the alignment results. *De novo* gene prediction was performed using GENSCAN(Burge and Karlin 1997), AUGUSTUS(Stanke et al. 2006), and GLIMMERHMM(Majoros et al. 2004) based on the repeat-masked genome. Then, EVM(Haas et al. 2008) and MAKER(Cantarel et al. 2008) were applied to integrate the predicted genes. Finally, manual integration was performed to construct the final gene set. We searched the final gene set against the KEGG(Kanehisa and Goto 2000), SwissProt(Bairoch and Apweiler 2000), and TrEMBL(Bairoch and Apweiler 2000) protein databases to identify gene functions. The gene motifs and domains were determined using InterProScan(Zdobnov and Apweiler 2001) following analysis of public protein databases, including ProDom, PRINTS, PFAM, SMART, PANTHER and PROSITE. All genes were aligned against the KEGG pathway database(Kanehisa and Goto 2000), and the best match for each gene was identified. The GO IDs for each gene were obtained from the corresponding InterPro entries. We also mapped milu proteins to the NCBI nr database and retrieved GO IDs using BLAST2GO(Conesa et al. 2005).

Genome evolution

Orthologous groups were constructed by ORTHOMCL v2.0.9. Phylogenetic tree inference and divergence time estimation was conducted based on fourfold-degenerate sites of single-copy gene families. Significantly expanded and contracted gene families were identified by CAFE(De Bie et al. 2006). Molecular evolution analyses were performed using the framework provided by the PAML4.7 package. Please see Supplementary information for more detailed methodologies.

Detection of variants

For the individual that was used for *de novo* sequencing, we used the BWA(Li and Durbin 2009) program to remap the pair-end (180 bp, 500 bp, and 800 bp) clean reads to the assembled scaffolds. After merging the BWA results and sorting alignments (using the leftmost coordinates) and removing potential PCR duplicates, we used SAMtools(Li et al. 2009) mpileup to call SNPs and short InDels. We applied vcfutils.pl varFilter (in SAMtools) as the filtering tool with parameters '*-Q 20 -d 6 -D 86*'. Then, homologous SNP positions were extracted and further filtered, to disqualify SNPs that may have resulted from errors due to assembly and/or mapping. The heterozygosity rate was estimated as the density of heterozygous SNPs for the whole genome, gene intervals, introns, and exons, respectively. For the five resequencing milu individuals, variants were identified using similar methods, except that the filtering parameter used by vcfutils.pl varFilter was '*-Q 20 -d 6 -D 75*'.

Whole genome re-sequencing data from 34 giant panda genomes(Zhao et al. 2013), and eight crested ibis(Li

et al. 2014) genomes were downloaded from the NCBI SRA database, and BAM files were generated using identical methods to those used for milu individuals. Next, the bam files for each species were processed using the mpileup module in samtools and the following parameters; '-q 1 -C 50 -g -t DP,SP,DP4 -I -d 250 -L 250 -m 2 -p'. The associated variants were called and filtered using the varFilter module of vcfutils.pl (parameters '*-Q 20 -d 10 -D 50000 –w 5 -W 10*' for panda, and '-Q 20 -d 5 -D 4000 -w 5 -W 10' for crested ibis). Finally, variants from each individual were generated by filtering positions with low depth ('<3' for panda, and '<5' for crested ibis). The SNP positions in 18 polar bear genomes(Liu et al. 2014) were extracted from variant files downloaded from GigaDB(Sneddon et al. 2012). SNPs were annotated using snpEff (Cingolani et al. 2012). To estimate how the functional changes for proteins in milu/panda/polar bear/crested ibis differed from those in humans, we evaluated the likely effect of a mutation in humans relative to the milu/panda/polar bear/crested ibis alleles as either neutral or deleterious using SIFT(Ng and Henikoff 2003).

Demographic history reconstruction and ROH identification

Demographic histories of the milu were reconstructed using the Pairwise Sequentially Markovian Coalescent (PSMC) model(Li and Durbin 2011). The mutation rate (μ) was set to 1.5×10^{-8} and the generation time (g) was set to 6 years. We identified the ROH for each individual using the runs of homozygosity tool in PLINK (v.1.07)(Purcell et al. 2007) with adjusted parameters (--homozyg-window-kb 0 --homozyg-window-snp 65 --homozyg-window-het 1 --homozyg-window-missing 3 --homozyg-window-threshold 0.05 --homozyg-snp 65 --homozyg-kb 100 --homozyg-density 5000 --homozyg-gap 5000). The individual genome-based inbreeding coefficient, denoted as *Froh*, is defined as the fraction of total ROH length to genome effective length(Gazal et al. 2014).

SNP densities

To check the distribution pattern of SNPs in the genomes, we adopted a method that was described by Hacquard *et al.*(Hacquard et al. 2013) Specifically, to estimate the distributions of the high- and low-SNP densities, we fitted a two-component mixture model to the observed SNP densities using the expectation-maximization (EM) algorithm (function normalmixEM, R-package mixtools). SNP densities were obtained via a sliding window of 200 kb, at steps of 2 kb, in scaffolds with lengths longer than 300kb. To identify regions with high- and low-SNP densities, a two-state hidden Markov model (HMM) was fitted on the 200-kb SNP densities using the EM algorithm, and the posterior state sequence was computed via the Viterbi algorithm (function fit, package depmixS4).

Selective sweep identification

To detect putative selective sweeps, we searched genomic regions with higher degrees of fixation, following previously described methods(Rubin et al. 2010; Rubin et al. 2012). The numbers of major and minor allele reads observed at each variant position were counted, and SNP positions which located on non-autosomes and whose minor allele frequency was <0.05 were filtered. We then scanned the genome using sliding 100-kb windows with a step size of 50 kb. Windows with less than five SNPs were not considered. Windows with ZHp ≤–6 were retained as candidate selective sweeps.

Salinity analyses

5 mg, 10 mg, 15 mg, 20 mg, 25 mg, 40 mg, 60 mg, 80 mg, 100 mg, 120 mg, 140 mg, 160 mg, 180 mg, and 200 mg of NaCl were weighed respectively in separate beakers. A total of 50 ml of distilled water was subsequently mixed with each quantity of NaCl to prepare saline standards. The electric conductivity (EC) value of standard saline was determined using a conductivity meter and the resultant values were used to generate the X-axis. The saline standard concentration values were used as the Y-axis. A total of 0.5 g of plant materials was weighed in a beaker and mixed with 100 ml of distilled water. After the mixture was heated using an electric stove for 30 min, the resultant solution was strained into a new volumetric flask with 25 ml of distilled water. The solution was stored in a 50-milliliter centrifuge tube and was subsequently used to determine EC values.

Metagenomics analyses

10 fresh fecal samples from three core areas in Dafeng Natural Reserve (China) were collected immediately after defecation, snap-frozen in liquid N2, and shipped to the laboratory on dry ice. All samples were obtained from inside the feces, where there was no contact with soil. DNA was extracted from fecal samples using the Qiagen QIAamp DNA Stool Mini Kit according to the protocol for isolation of DNA for pathogen detection. DNA was eluted in a final volume of 250 μL using elution buffer and then stored at −20 °C. Sequencing and general data analyses were performed by Shanghai Majorbio Bio-pharm Biotechnology (Shanghai, China). A library was constructed with an average clone insert size of 350-bp for each sample. We compared the raw short reads with host genome data to remove the host sequence. Clean reads were subsequently obtained to assemble long contig sequences using SOAPdenovo(Li et al. 2010) during metagenomic analyses. Different Kmer frequencies were utilized to generate different assembly results, and N50 lengths were used to access the best assembly result. The metagenomes were uploaded to MG-RAST. Functional annotation of 49 metagenomes (10 from milu and 39 from published data) was performed with Hierarchical Classification using the KEGG ortholog database within MG-RAST(Meyer et al. 2008). The following parameters were used: maximum e-value cutoff of 1e-5, minimum

identity cutoff of 60%, and minimum alignment length cutoff of 15 (default). The statistical analysis for KEGG function pathways were performed in STAMP(Parks et al. 2014).

Reference

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of molecular biology* **215**(3): 403-410.

Bairoch A, Apweiler R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic acids research* **28**(1): 45-48.

Bakewell MA, Shi P, Zhang J. 2007. More genes underwent positive selection in chimpanzee evolution than in human evolution. *Proceedings of the National Academy of Sciences* **104**(18): 7489-7494.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research* **27**(2): 573.

Birney E, Clamp M, Durbin R. 2004. GeneWise and genomewise. *Genome research* **14**(5): 988-995.

Boehm U, Klamp T, Groot M, Howard JC. 1997. Cellular responses to interferon-gamma. *Annual review of immunology* **15**: 749-795.

Bouchard MJ, Dong Y, McDermott BM, Jr., Lam DH, Brown KR, Shelanski M, Bellve AR, Racaniello VR. 2000. Defects in nuclear and cytoskeletal morphology and mitochondrial localization in spermatozoa of mice lacking nectin-2, a component of cell-cell adherens junctions. *Molecular and cellular biology* **20**(8): 2865-2873.

Burge C, Karlin S. 1997. Prediction of complete gene structures in human genomic DNA. *Journal of molecular biology* **268**(1): 78-94.

Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Alvarado AS, Yandell M. 2008. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome research* **18**(1): 188-196.

Cao K. 2005. *Research on the Mi-deer*. Shanghai Publishing House for the Science and Technology Education, Shanghai, China.

Chandrashekar J, Kuhn C, Oka Y, Yarmolinsky DA, Hummler E, Ryba NJ, Zuker CS. 2010. The cells and peripheral representation of sodium taste in mice. *Nature* **464**(7286): 297-301.

Choi H-S, Lee S-H, Kim H, Lee Y. 2008. Germ cell-specific gene 1 targets testis-specific poly (A) polymerase to the endoplasmic reticulum through protein–protein interactions. *FEBS letters* **582**(8): 1203-1209.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 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* **6**(2): 80-92.

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**(18): 3674-3676.

Curik I, Ferenčaković M, Sölkner J. 2014. Inbreeding and runs of homozygosity: a possible solution to an old problem. *Livestock Science* **166**: 26-34.

De Bie T, Cristianini N, Demuth JP, Hahn MW. 2006. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* **22**(10): 1269-1271.

Deshpande DA, Wang WC, McIlmoyle EL, Robinett KS, Schillinger RM, An SS, Sham JS, Liggett SB. 2010. Bitter taste receptors on airway smooth muscle bronchodilate by localized calcium signaling and reverse obstruction.

Nature medicine **16**(11): 1299-1304.

Frankham R. 2005. Genetics and extinction. *Biological conservation* **126**(2): 131-140.

Gazal S, Sahbatou M, Perdry H, Letort S, Génin E, Leutenegger A-L. 2014. Inbreeding coefficient estimation with dense SNP data: comparison of strategies and application to HapMap III. *Human heredity* **77**(1-4): 49-62.

Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome biology* **9**(1): R7.

Hacquard S, Kracher B, Maekawa T, Vernaldi S, Schulze-Lefert P, van Themaat EVL. 2013. Mosaic genome structure of the barley powdery mildew pathogen and conservation of transcriptional programs in divergent hosts. *Proceedings of the National Academy of Sciences* **110**(24): E2219-E2228.

Holstein AF. 1976. Ultrastructural observations on the differentiation of spermatids in man. *Andrologia* **8**(2): 157-165.

Jiang Z, Harris RB. 2008. *Elaphurus davidianus*. In *The IUCN Red List of Threatened Species*, Vol 23 July 2014.

Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28**(1): 27-30.

Keller LF, Waller DM. 2002. Inbreeding effects in wild populations. *Trends in Ecology & Evolution* **17**(5): 230-241.

Kim E-S, Cole JB, Huson H, Wiggans GR, Van Tassell CP, Crooker BA, Liu G, Da Y, Sonstegard TS. 2013. Effect of artificial selection on runs of homozygosity in US Holstein cattle. *PLoS One* **8**(11): e80813.

Ley RE, Hamady M, Lozupone C, Turnbaugh PJ, Ramey RR, Bircher JS, Schlegel ML, Tucker TA, Schrenzel MD, Knight R et al. 2008. Evolution of mammals and their gut microbes. *Science* **320**(5883): 1647-1651.

Li C, Yang X, Ding Y, Zhang L, Fang H, Tang S, Jiang Z. 2011. Do Père David's Deer Lose Memories of Their Ancestral Predators? *PLoS ONE* **6**(8): e23623.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**(14): 1754-1760.

-. 2011. Inference of human population history from individual whole-genome sequences. *Nature* **475**(7357): 493-496.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**(16): 2078-2079.

Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K. 2010. De novo assembly of human genomes with massively parallel short read sequencing. *Genome research* **20**(2): 265-272.

Li S, Li B, Cheng C, Xiong Z, Liu Q, Lai J, Carey HV, Zhang Q, Zheng H, Wei S. 2014. Genomic signatures of near-extinction and rebirth of the crested ibis and other endangered bird species. *Genome biology* **15**(12): 1-17.

Liu S, Lorenzen ED, Fumagalli M, Li B, Harris K, Xiong Z, Zhou L, Korneliussen TS, Somel M, Babbitt C. 2014. Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears. *Cell* **157**(4): 785-794.

Majoros WH, Pertea M, Salzberg SL. 2004. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**(16): 2878-2879.

Malenfant RM, Davis CS, Moore SS, Coltman DW. 2014. White-tailed deer (Odocoileus virginianus) transcriptome assembly and SNP discovery *Molecular Ecology Resources* **accepted**.

Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A et al. 2008. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC bioinformatics* **9**: 386.

Miki K, Qu W, Goulding EH, Willis WD, Bunch DO, Strader LF, Perreault SD, Eddy EM, O'Brien DA. 2004. Glyceraldehyde 3-phosphate dehydrogenase-S, a sperm-specific glycolytic enzyme, is required for sperm motility and male fertility. *Proceedings of the National Academy of Sciences of the United States of America* **101**(47): 16501-16506.

Muegge BD, Kuczynski J, Knights D, Clemente JC, Gonzalez A, Fontana L, Henrissat B, Knight R, Gordon JI. 2011. Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science*

332(6032): 970-974.

Ng PC, Henikoff S. 2003. SIFT: Predicting amino acid changes that affect protein function. *Nucleic acids research* **31**(13): 3812-3814.

Parks DH, Tyson GW, Hugenholtz P, Beiko RG. 2014. STAMP: statistical analysis of taxonomic and functional profiles. *Bioinformatics* **30**(21): 3123-3124.

Parra G, Bradnam K, Korf I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**(9): 1061-1067.

Price AL, Jones NC, Pevzner PA. 2005. De novo identification of repeat families in large genomes. *Bioinformatics* **21**(suppl 1): i351-i358.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* **81**(3): 559-575.

Qiu Q, Zhang G, Ma T, Qian W, Wang J, Ye Z, Cao C, Hu Q, Kim J, Larkin DM. 2012. The yak genome and adaptation to life at high altitude. *Nature genetics* **44**(8): 946-949.

Quill TA, Sugden SA, Rossi KL, Doolittle LK, Hammer RE, Garbers DL. 2003. Hyperactivated sperm motility driven by CatSper2 is required for fertilization. *Proceedings of the National Academy of Sciences* **100**(25): 14869-14874.

Rubin C-J, Megens H-J, Barrio AM, Maqbool K, Sayyab S, Schwochow D, Wang C, Carlborg Ö, Jern P, Jørgensen CB. 2012. Strong signatures of selection in the domestic pig genome. *Proceedings of the National Academy of Sciences* **109**(48): 19529-19536.

Rubin C-J, Zody MC, Eriksson J, Meadows JR, Sherwood E, Webster MT, Jiang L, Ingman M, Sharpe T, Ka S. 2010. Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* **464**(7288): 587-591.

Saccheri I, Kuussaari M, Kankare M, Vikman P, Fortelius W, Hanski I. 1998. Inbreeding and extinction in a butterfly metapopulation. *Nature* **392**(6675): 491-494.

Sanders JG, Beichman AC, Roman J, Scott JJ, Emerson D, McCarthy JJ, Girguis PR. 2015. Baleen whales host a unique gut microbiome with similarities to both carnivores and herbivores. *Nature communications* **6**.

Sequencing TC, Consortium A. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**(7055): 69-87.

Sheppard DN, Welsh MJ. 1999. Structure and function of the CFTR chloride channel. *Physiological reviews* **79**(1 Suppl): S23-45.

Smit AF, Hubley R, Green P. 1996. RepeatMasker Open-3.0.

Sneddon TP, Li P, Edmunds SC. 2012. GigaDB: announcing the GigaScience database. *GigaScience* **1**(1): 11.

Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. 2006. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic acids research* **34**(suppl 2): W435-W439.

Steiner CC, Putnam AS, Hoeck PE, Ryder OA. 2013. Conservation genomics of threatened animal species. *Annu Rev Anim Biosci* **1**(1): 261-281.

Wan Q-H, Pan S-K, Hu L, Zhu Y, Xu P-W, Xia J-Q, Chen H, He G-Y, He J, Ni X-W. 2013. Genome analysis and signature discovery for diving and sensory properties of the endangered Chinese alligator. *Cell research* **23**(9): 1091-1105.

Xu Z, Wang H. 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic acids research* **35**(suppl 2): W265-W268.

Xue Y, Prado-Martinez J, Sudmant PH, Narasimhan V, Ayub Q, Szpak M, Frandsen P, Chen Y, Yngvadottir B, Cooper DN. 2015. Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. *Science* **348**(6231): 242-245.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution* **24**(8): 1586-1591.

Yao B, Zhao Y, Wang Q, Zhang M, Liu M, Liu H, Li J. 2012a. De novo characterization of the antler tip of Chinese Sika deer transcriptome and analysis of gene expression related to rapid growth. *Molecular and cellular biochemistry* **364**(1-2): 93-100.

Yao B, Zhao Y, Zhang H, Zhang M, Liu M, Liu H, Li J. 2012b. Sequencing and de novo analysis of the Chinese Sika deer antler-tip transcriptome during the ossification stage using Illumina RNA-Seq technology. *Biotechnology letters* **34**(5): 813-822.

Yeyati PL, Bancewicz RM, Maule J, van Heyningen V. 2007. Hsp90 selectively modulates phenotype in vertebrate development. *PLoS Genet* **3**(3): e43.

Yim H-S, Cho YS, Guang X, Kang SG, Jeong J-Y, Cha S-S, Oh H-M, Lee J-H, Yang EC, Kwon KK. 2014. Minke whale genome and aquatic adaptation in cetaceans. *Nature genetics* **46**(1): 88-92.

Yokoyama Y, Lambeck K, De Deckker P, Johnston P, Fifield LK. 2000. Timing of the Last Glacial Maximum from observed sea-level minima. *Nature* **406**(6797): 713-716.

Zdobnov EM, Apweiler R. 2001. InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**(9): 847-848.

Zhao S, Zheng P, Dong S, Zhan X, Wu Q, Guo X, Hu Y, He W, Zhang S, Fan W. 2013. Whole-genome sequencing of giant pandas provides insights into demographic history and local adaptation. *Nature Genetics* **45**(1): 67-71.

Zhou X, Sun F, Xu S, Fan G, Zhu K, Liu X, Chen Y, Shi C, Yang Y, Huang Z. 2013. Baiji genomes reveal low genetic variability and new insights into secondary aquatic adaptations. *Nature communications* **4**.

Zhou X, Wang B, Pan Q, Zhang J, Kumar S, Sun X, Liu Z, Pan H, Lin Y, Liu G. 2014. Whole-genome sequencing of the snub-nosed monkey provides insights into folivory and evolutionary history. *Nature genetics* **46**(12): 1303-1310.