**Supplemental Material**

# GC-biased gene conversion drives accelerated evolution of ultraconserved elements in mammalian and avian genomes

Anguo Liu[#], Nini Wang[#], Guoxiang Xie[#], Yang Li, Xixi Yan, Xinmei Li, Zhenliang Zhu, Zhuohui Li, Jing Yang, Fanxin Meng, Mingle Dou, Weihuang Chen, Nange Ma, Yu Jiang, Yuanpeng Gao[*], Yu Wang[*]

[#] These authors contributed equally to this work.

*Corresponding author:

Yu Wang, wang_yu@nwafu.edu.cn

Yuanpeng Gao, gaoyuanpeng1990@163.com

**This PDF file includes:**

Supplemental Methods and Results

Supplemental Figures S1-S17

Supplemental Tables S1-S21 legends

Supplemental Data S1-S4 legends

**The materials presented as separate files include:**

Supplemental Tables S1-S21 (in Supplemental_Tables.xlsx)

Supplementary Data S1-S4 (in Supplemental_Data.zip)

Supplementary Code (in Supplemental_Code.zip)

## Supplemental Methods and Results

### Analysis of 7 non-overlapped human-rodent UCEs

For the remaining 7 human-rodent UCEs that did not overlap with our candidate mUCEs, we examined and visualized them in the Genome Browser (Gonzalez et al. 2021) to identify the possible reasons. We found that the exclusion of these seven UCEs from our candidate mUCE set can be summarized in three reasons: (1) a lack of orthologous sequences (uc.64, uc.125, uc.453, and uc.454) in the reconstructed ancestral genome due to the relatively low quality of the reference genome (armadillo), (2) imperfect ultra-conservation (uc.415 and uc.476) in most lineages, and (3) a one-to-many orthology relationship (uc.471) in the human/mouse/rat genome.

3 out of 4 missing orthologous sequences can be found in the recently released high-quality armadillo assembly (GCF_030445035.1) using BLAST v2.11.0 (Camacho et al. 2009) with the parameter settings: "-task blastn, -outfmt 6, -max_target_seqs 1, -max_hsps 1", except for uc.454, which has excessive substitutions in the armadillo genome. Although it is expected that using more complete genome assemblies would identify more UCEs in both taxa, our final mUCEs and aUCEs are accurate and thus would not affect our analyses.

### Investigation of gene completeness of ZNF536

Using pairwise alignments of *ZNF536* coding sequence (conserved canonical transcript, ENST00000355537) generated by TOGA (https://genome.senckenberg.de/download/TOGA/human_hg38_reference/) (Kirilenko et al. 2023), we confirmed that *ZNF536* were complete in both Yingochiroptera and Yangochiroptera without any inactivating mutations at least at the family level. Notably, the *ZNF536* coding sequence in the same assemblies that we used in our analyses were classified as "missing", "partial intact" or "uncertain loss" by TOGA, where <50% central part of their *ZNF536* coding region is completely present in the assembly. This is potentially attributed to assembly gaps or fragmentation of the low-quality (low Contig N50) assemblies (Kirilenko et al. 2023). All mentioned data can be found and visualized at: https://genome.senckenberg.de/TOGA.mammals.html.

### Validation of one-to-one orthologous UCEs

Focusing on one-to-one orthologous UCEs is important because fast divergence is one of key characters to paralogous genes and the corresponding regulatory elements. To accurately identify and analyze the target UCEs (mUCE.1304, mUCE.1639, mUCE.2036), our investigations were first based on synteny filter to ensure a conserved genomic context of target genes. Additionally, we manually checked the UCE-adjacent genes in Genome browser (Gonzalez et al. 2021) to confirm the presence of only one ortholog in query genomes.

To further exclude any possible paralogs of the three target UCEs presented in Figure 4-6 within each of the analyzed species, we conducted additional analyses using BLAT (Kent 2002) as follows:

1. For the case of mUCE.2036 in Chiroptera (Fig. 4), we obtained the corresponding UCE sequence in the hg38 assembly from our alignment. We then used BLAT (Kent 2002) to

search for all possible paralogos that mimic the real gBGC-induced fast-evolving UCEs in the query genomes (five bat genomes).

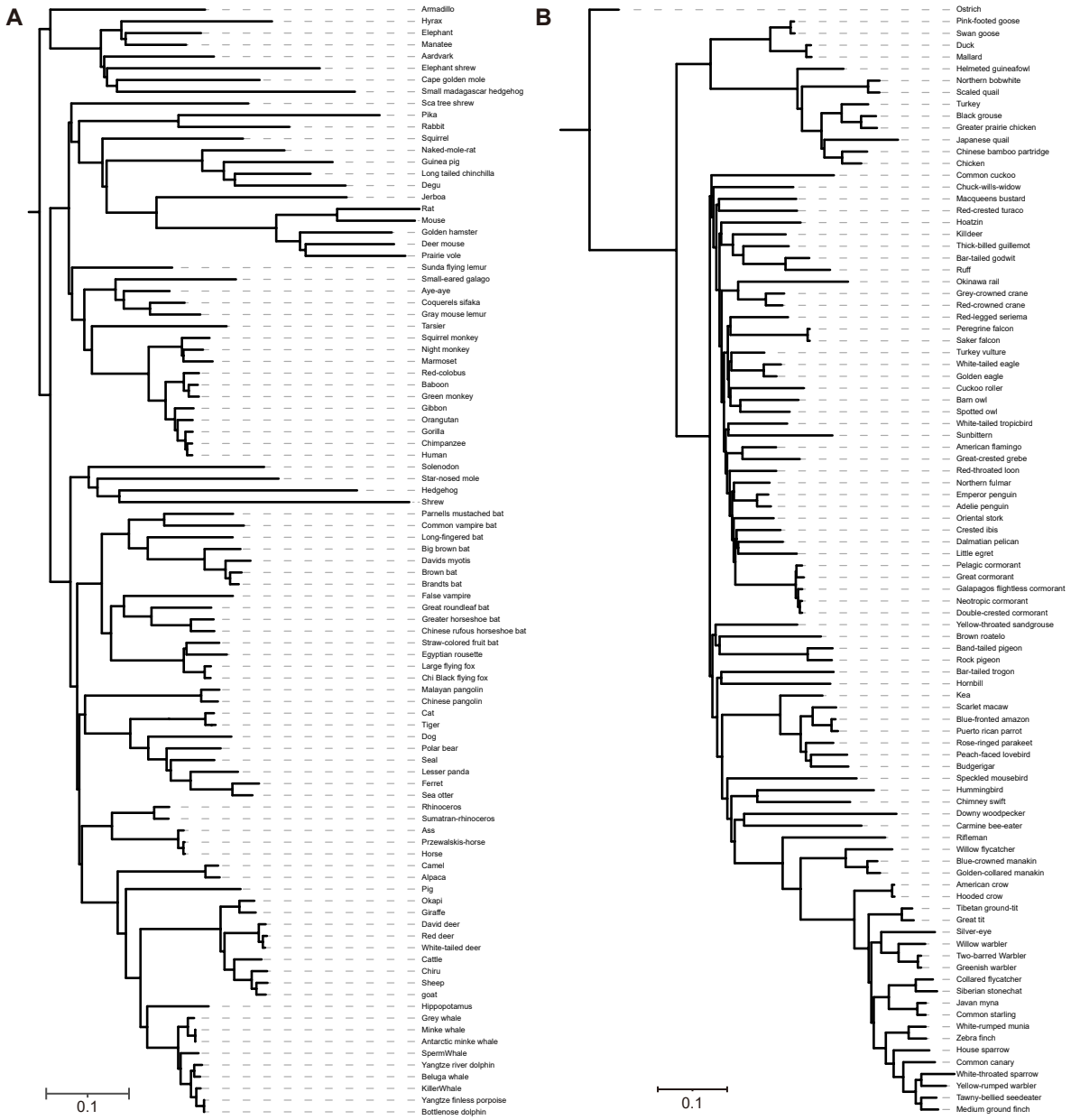
2. For the case of mUCE.1639 in Cervidae (Fig. 5), we obtained the corresponding UCE sequence in the hg38 assembly from our alignment. We then used BLAT (Kent 2002) to search for all possible paralogos that mimic the real gBGC-induced fast-evolving UCEs in the query genomes (three cervid genomes).
3. For the case of mUCE.1304 in placental mammals (Fig. 6), we obtained sequences of the corresponding avian UCE (aUCE.2913) from galGal6a assembly. We then used BLAT (Kent 2002) to identify all possible paralogos in human (hg38), rat (mm10), dog(canFam3), cattle (bosTau9) and platypus (ornAna2) assemblies.

The results showed that no paralogous sequences were identified by in all three cases. Furthermore, despite the significant lineage-specific acceleration observed in these UCEs (FDR < 0.05), they still exhibited high conservation and maintained their conserved genomic position.

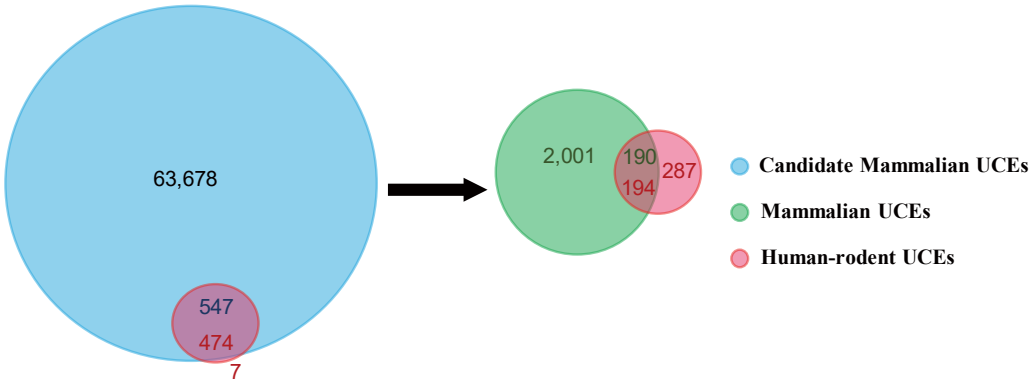
**Purifying selection analysis of cervid uc.359**

To further test whether the high conservation of uc.359 within cervids is a sign of purifying selection, we used 11-way ruminant alignment covering all six ruminant family (including six cervid species as same as shown in Supplemental Fig. S15) and estimated the neutral evolutionary rate using 4-fold degenerate sites. The 4-fold degenerate sites of all species were extracted from multiple sequence alignments according to the gene annotations of the reference genome.

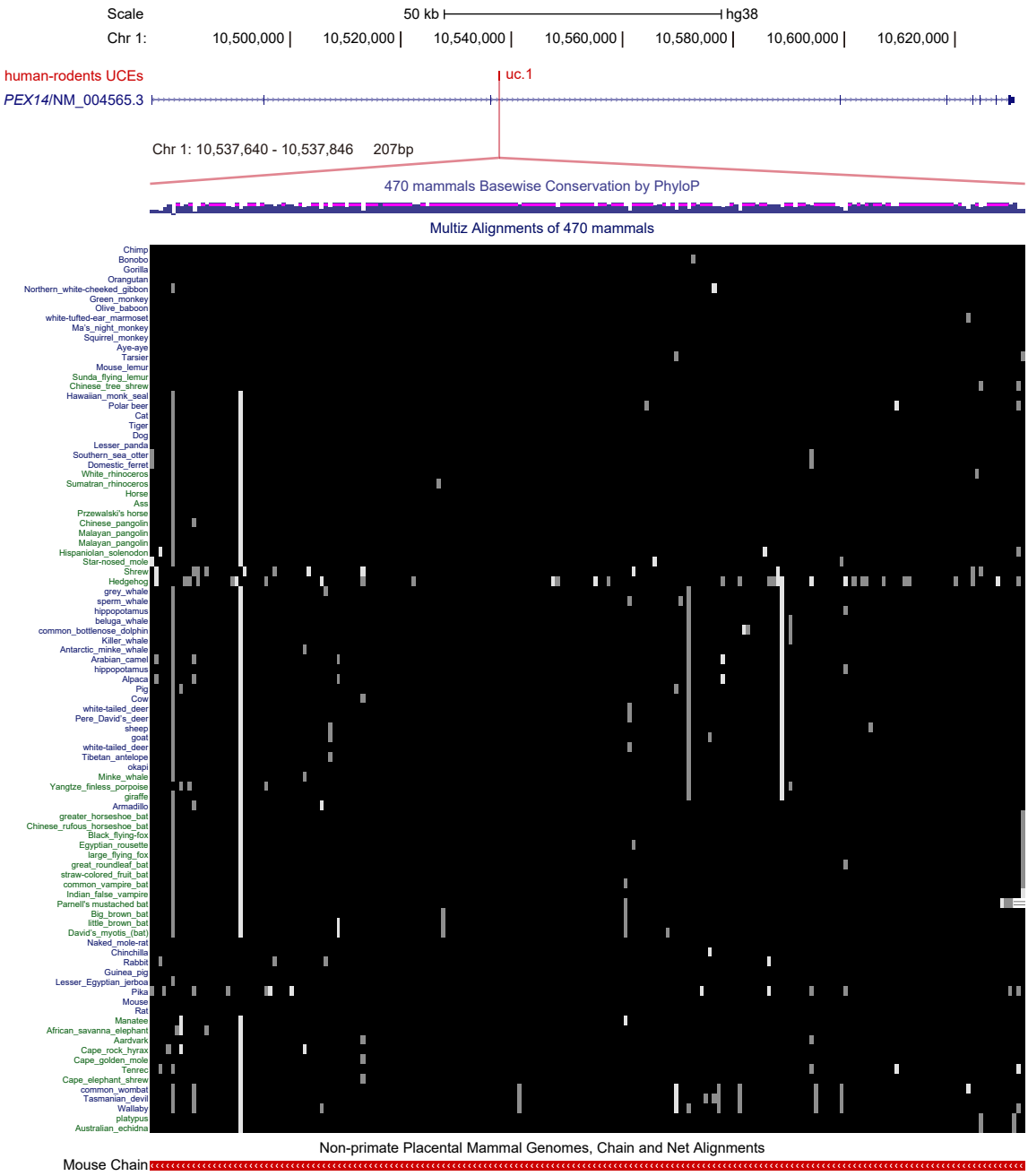
The alignment was extracted from the original alignment in our lab's website (http://animal.omics.pro/genomebrowser/cgi-bin/hgTablesCattle?clade=ruminantia&org=ARS_UCD_addY) (Fu et al. 2022). Detailed genome information can be found at: http://animal.omics.pro/code/index.php/RGD/loadByGet?address[]=RGD/Download/GenoDownload.php (Fu et al. 2022).



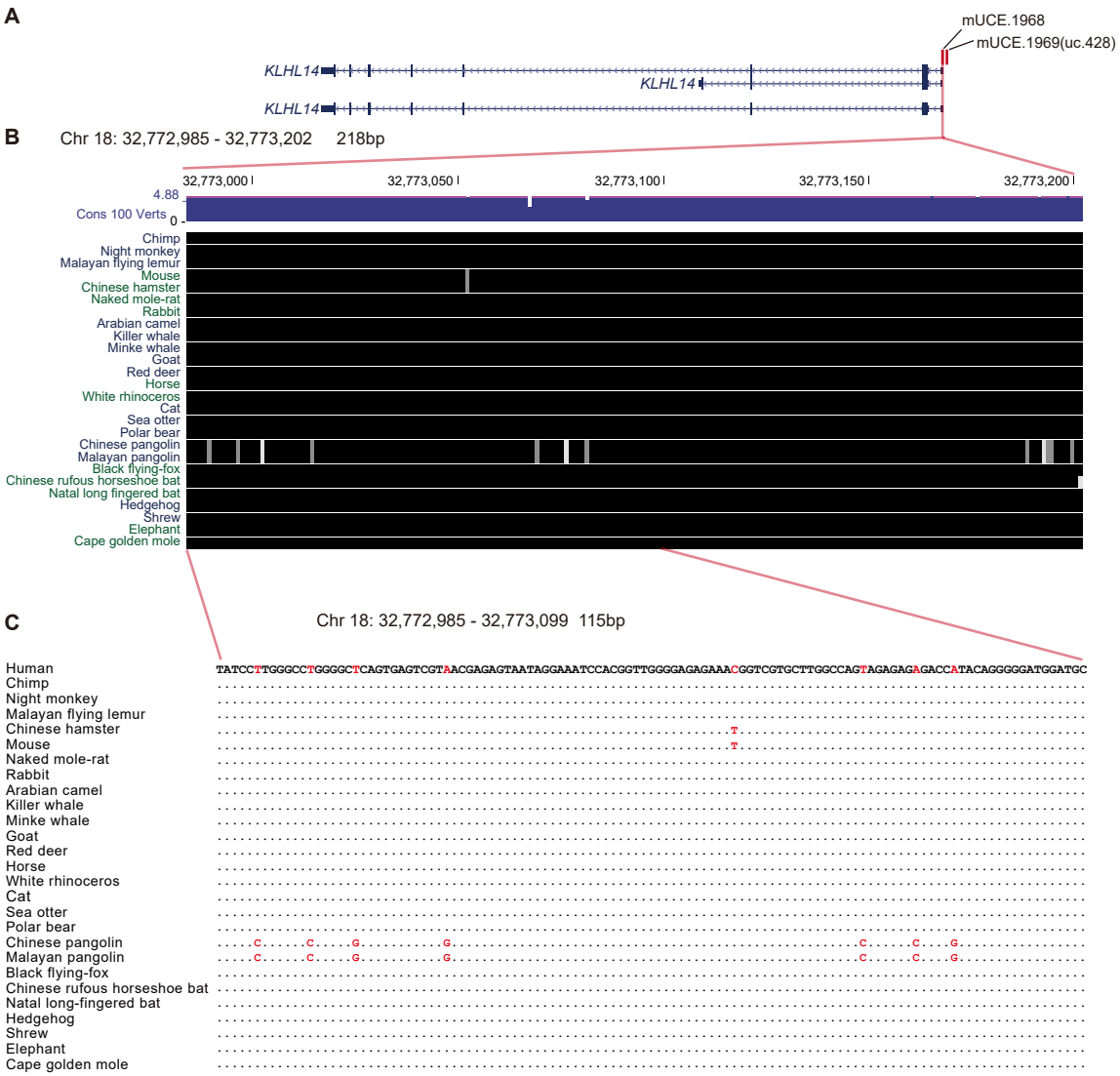

**Supplemental Fig. S1 Phylogenetic tree construction.** Phylogenomic constructions of (A) 95 mammals and (B) 94 birds inferred by GTR+F+I+G4 model in IQ-TREE (Nguyen et al. 2015) using 4-fold degenerate sites generated from whole genome alignment and the maximum likelihood method. Branch length presents the number of substitutions per site.



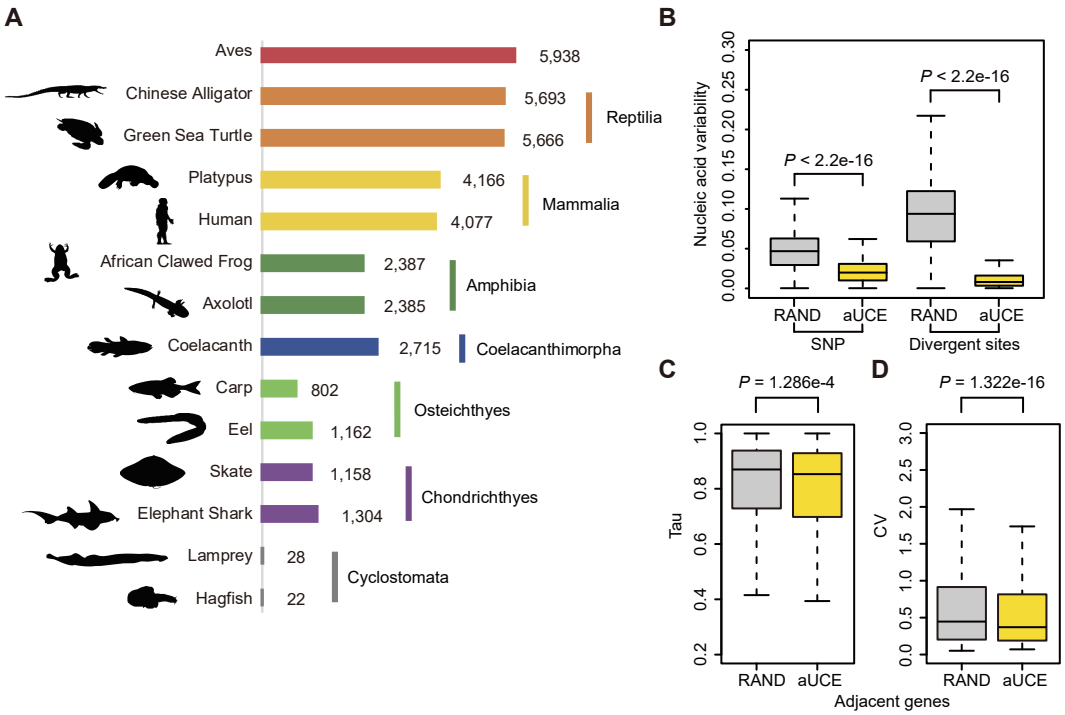

**Supplemental Fig. S2 Comparison of mUCEs in this study with 481 human-rodent UCEs.** There are 474 human-rodent UCEs still exist in the 64,678 candidate mUCEs in this study. 194 human-rodent UCEs are shared with the final 2,191 mUCEs we identified.



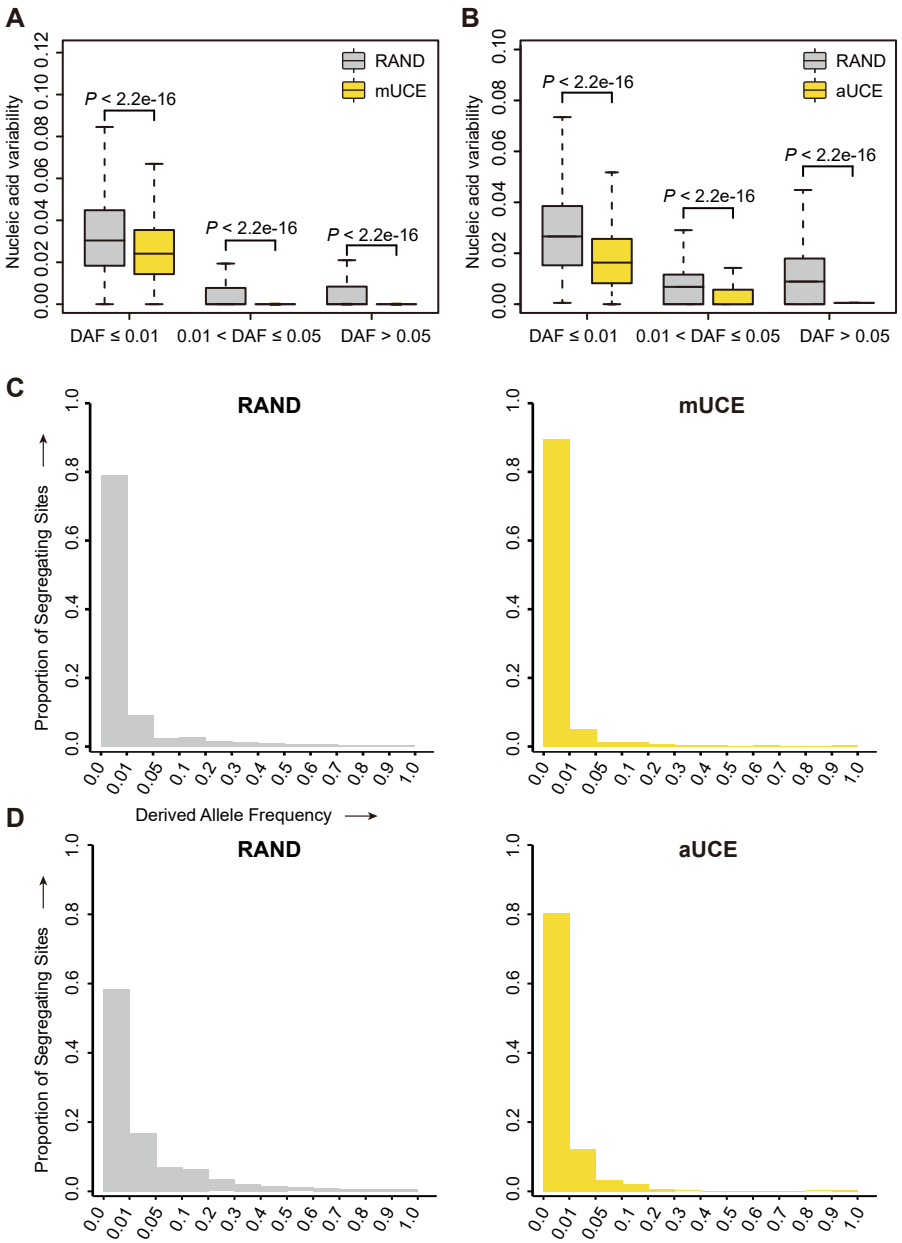

**Supplemental Fig. S3 Example of a human-rodents UCE excluded from our mammalian UCE set.** uc.1 is an example of human-rodents UCEs that was included in our candidate mUCEs but was subsequently filtered out due to excessive mutations. The 470-way alignment of uc.1 in the UCSC genome browser shows that uc.1 exhibits imperfect ultra-conservation across mammals and does not meet our two criteria.



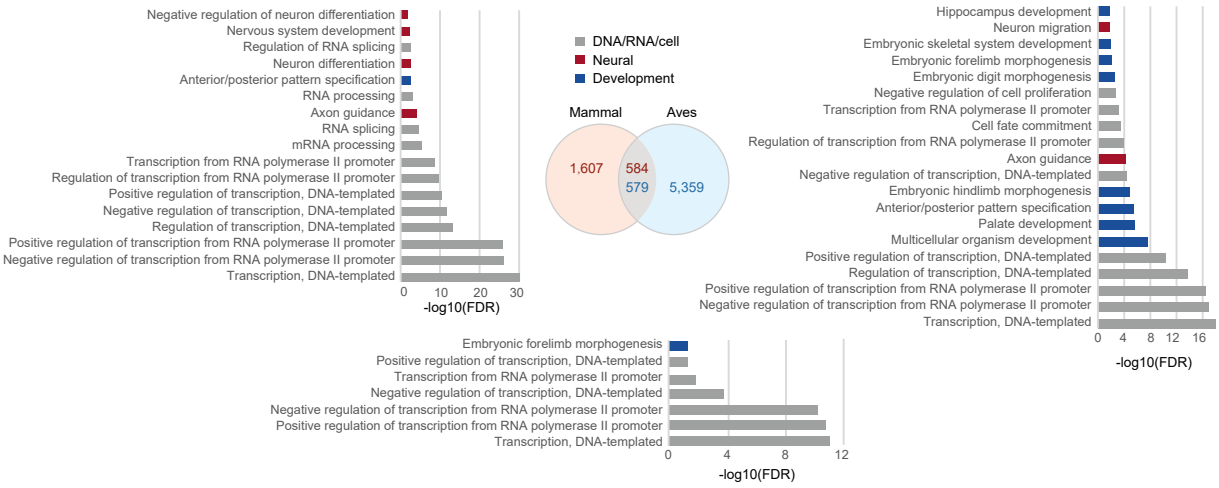

**Supplemental Fig. S4 Example of a new identified mammalian UCE with species-specific mutations.** (A) mUCE.1968 is located in the 3′UTR region of *KLHL14*. (B) Sequence alignment of 218-bp mUCE.1968 from UCSC genome browser. (C) Alignment of 115-bp sub-region shows that most sites are identical across mammals but also reveals a few pangolin-specific GC-biased substitutions. Dots in the sequence alignment refer to bases that are identical to those in the human genome. Substitutions are shown in red.



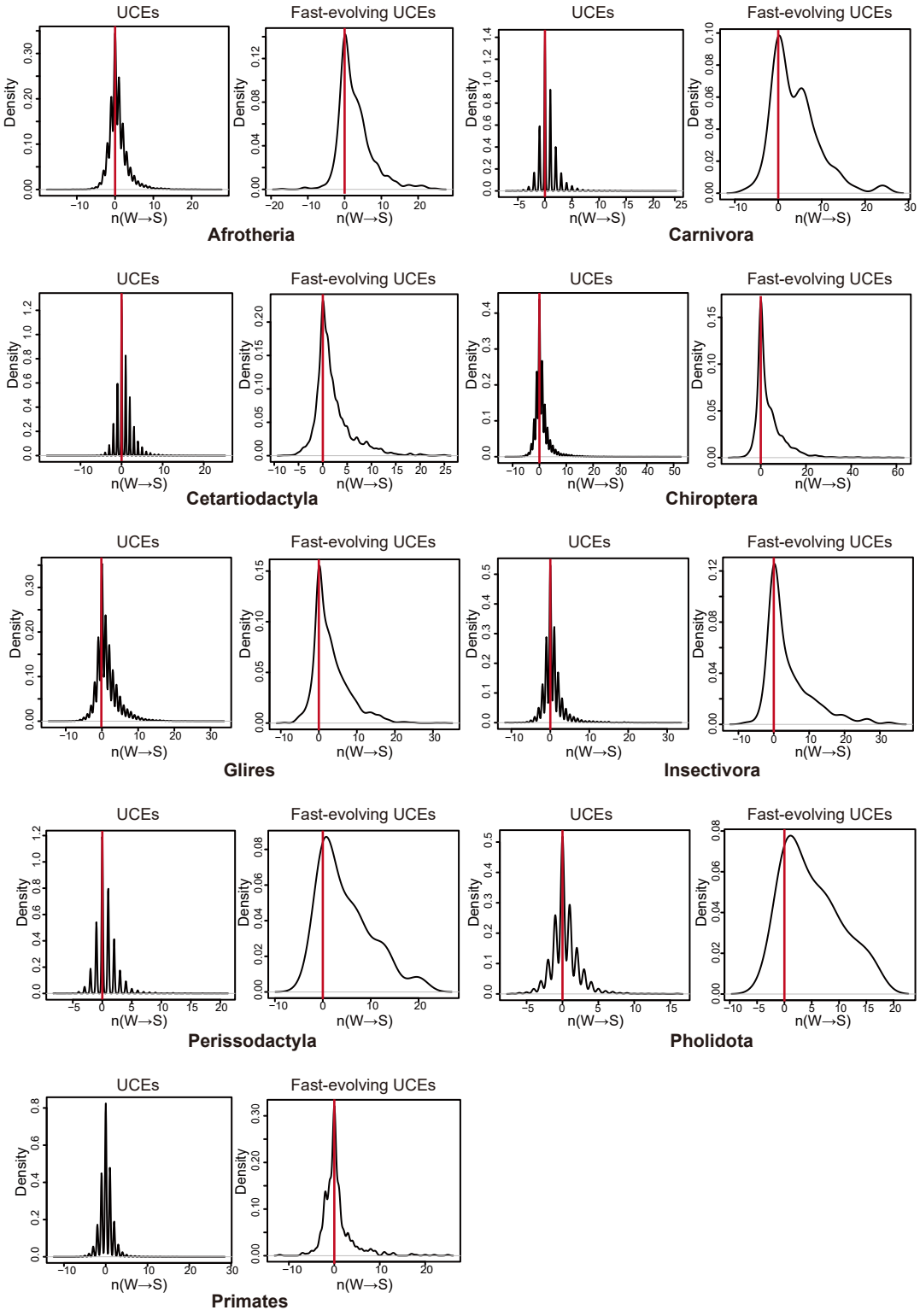

**Supplemental Fig. S5 Characteristics of Avian UCEs.** (A) Numbers of homologous sequences of avian UCEs (aUCEs) present in different vertebrate taxa. (B) Enrichment of chicken SNPs and counts of divergent sites between chickens and turkeys in aUCEs and RAND (randomly selected genomic regions). (C) Tissue-specific expression index (τ) of UCE-adjacent genes and RAND-adjacent genes (adjacent genes of randomly selected genomic regions). (D) Gene expression diversity value (CV) of UCE-adjacent genes and RAND-adjacent genes.




**Supplemental Fig. S6 Distribution signatures of SNPs in UCEs.** Enrichment of (A) human SNPs and (B) chicken SNPs in UCEs and RAND under different SNP frequency criteria is shown. Derived allele frequency (DAF) spectra representing (C) 3202 human individuals and (D) 928 chicken individuals for segregating sites within UCEs and RAND.

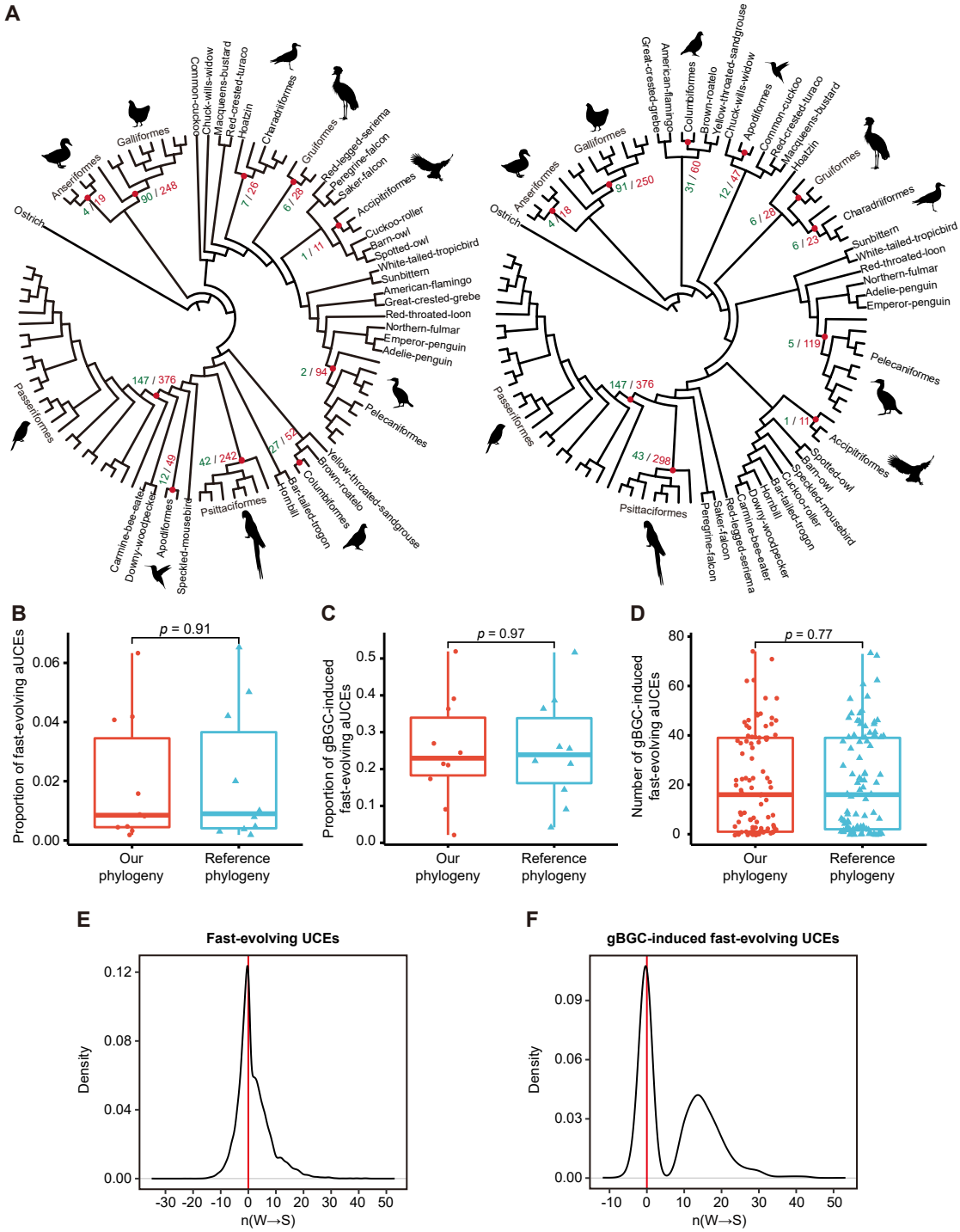



**Supplemental Fig. S7 Representative enriched Gene Ontology terms of mammalian and avian UCEs.** Representative GO enrichment results of adjacent genes of mUCEs (left), aUCEs (right) and they shared (below). GO terms related to regulation of DNA/RNA/cell, neural system and development are showed in gray, red and blue colors, respectively. The Venn chart indicates the number of mUCEs and aUCEs of which their adjacent genes were used to perform GO enrichment analysis.



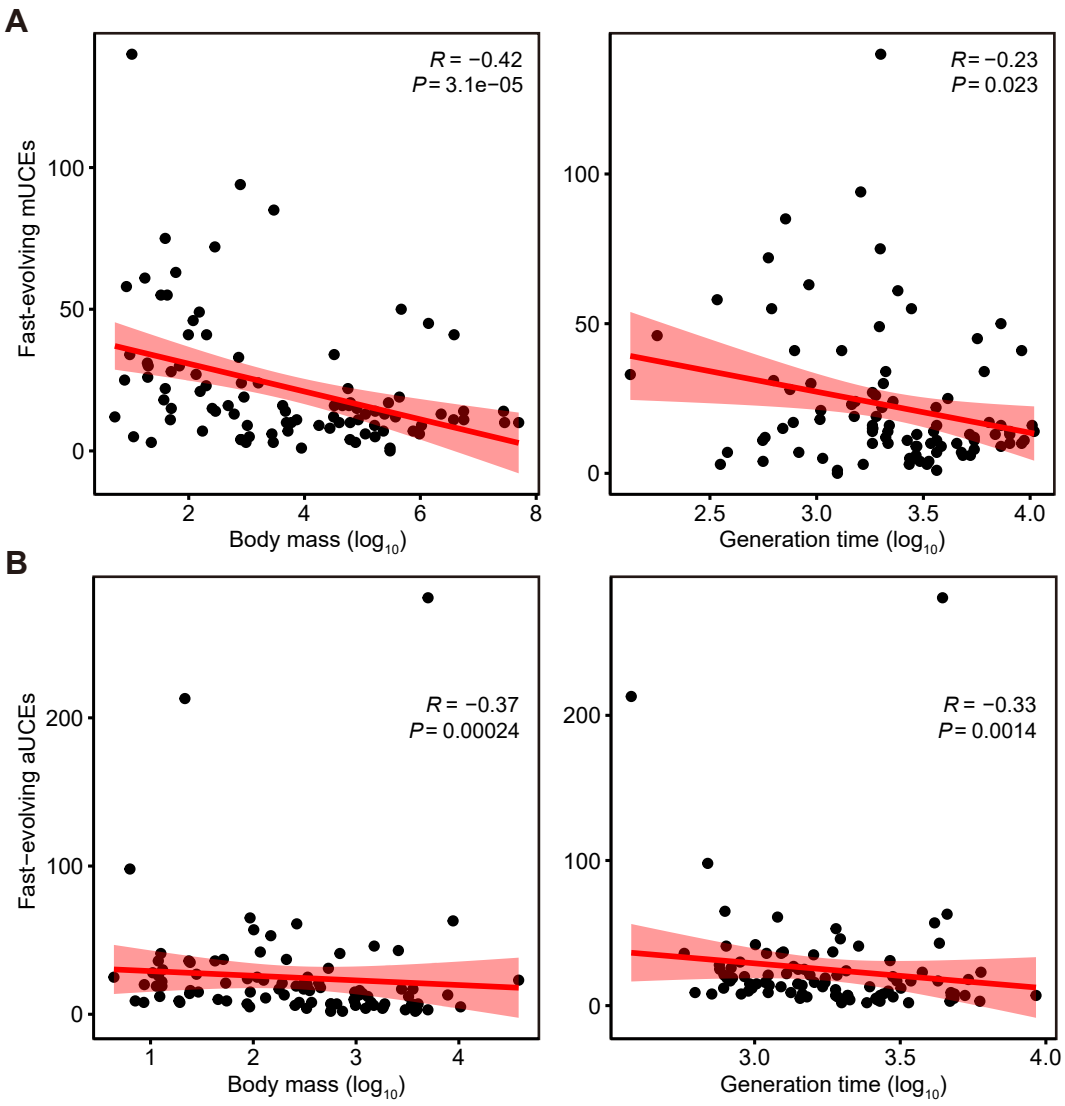

**Supplemental Fig. S8 gBGC trend among UCEs in specific mammalian lineages.**

Distribution of substitution numbers in UCEs and fast-evolving UCEs of all species in 9 mammalian lineages. The substitution numbers of each fast-evolving UCE and gBGC-induced fast-evolving UCE were counted according to the mutation direction, including S→W type and W→S type. The X axis indicates the number of W→S substitutions [n(W→S)]. S→W substitutions were scored as negative values.

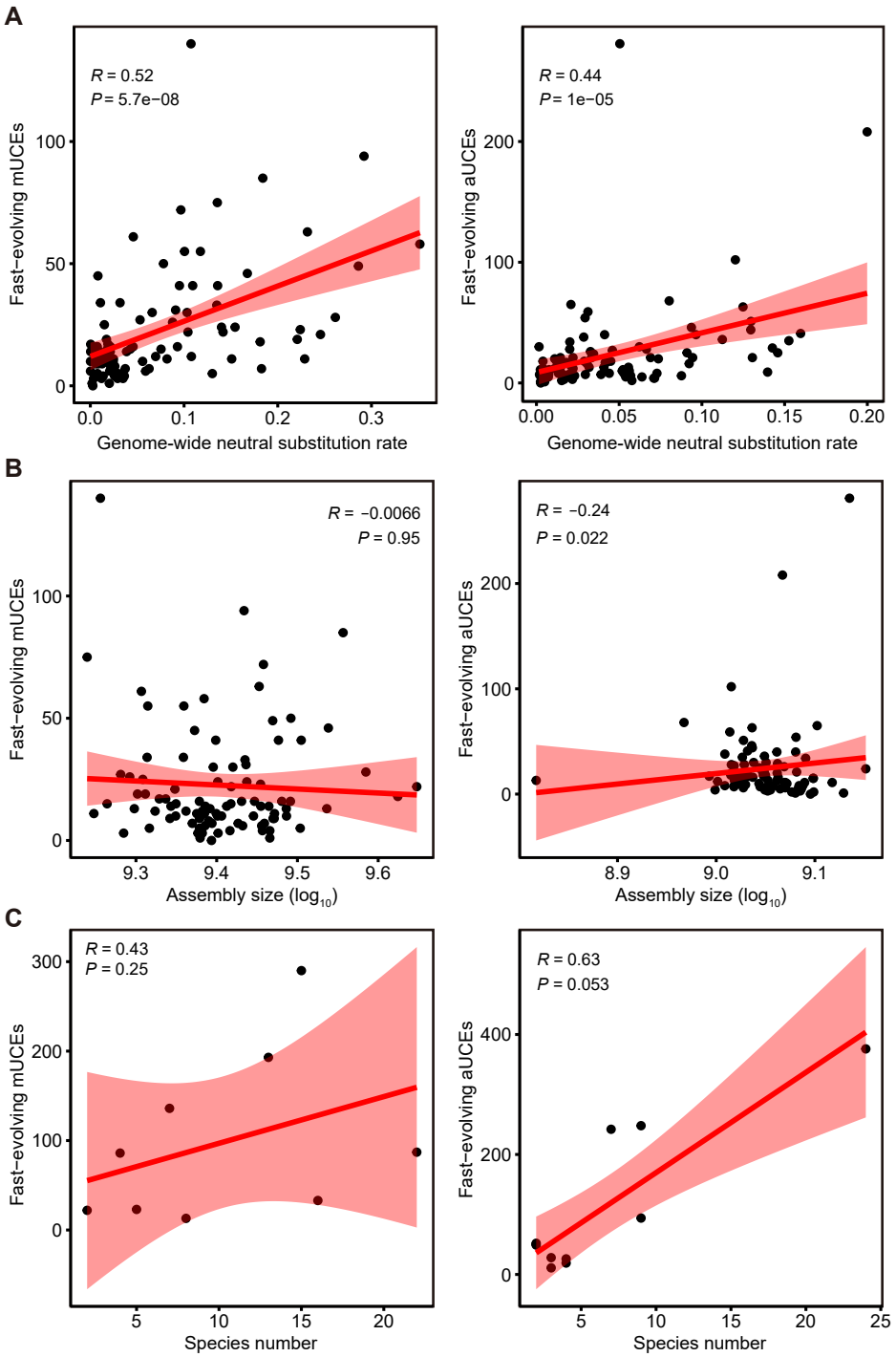



**Supplemental Fig. S9 Accelerated evolution and gBGC of UCEs in birds.** (A) Number of fast-evolving UCEs (red) and gBGC-induced fast-evolving UCEs (green) in specific avian lineages (red dots) based on our constructed phylogeny (left) and a reference phylogeny (right). (B) Proportion of fast-evolving UCEs in ten test avian lineages calculated based on two phylogeny. (C) Proportion of gBGC-induced fast-evolving UCEs in ten test avian lineages calculated based on two phylogeny. (D) Number of gBGC-induced fast-evolving UCEs in each avian species calculated based on two phylogeny. (E) Distribution of

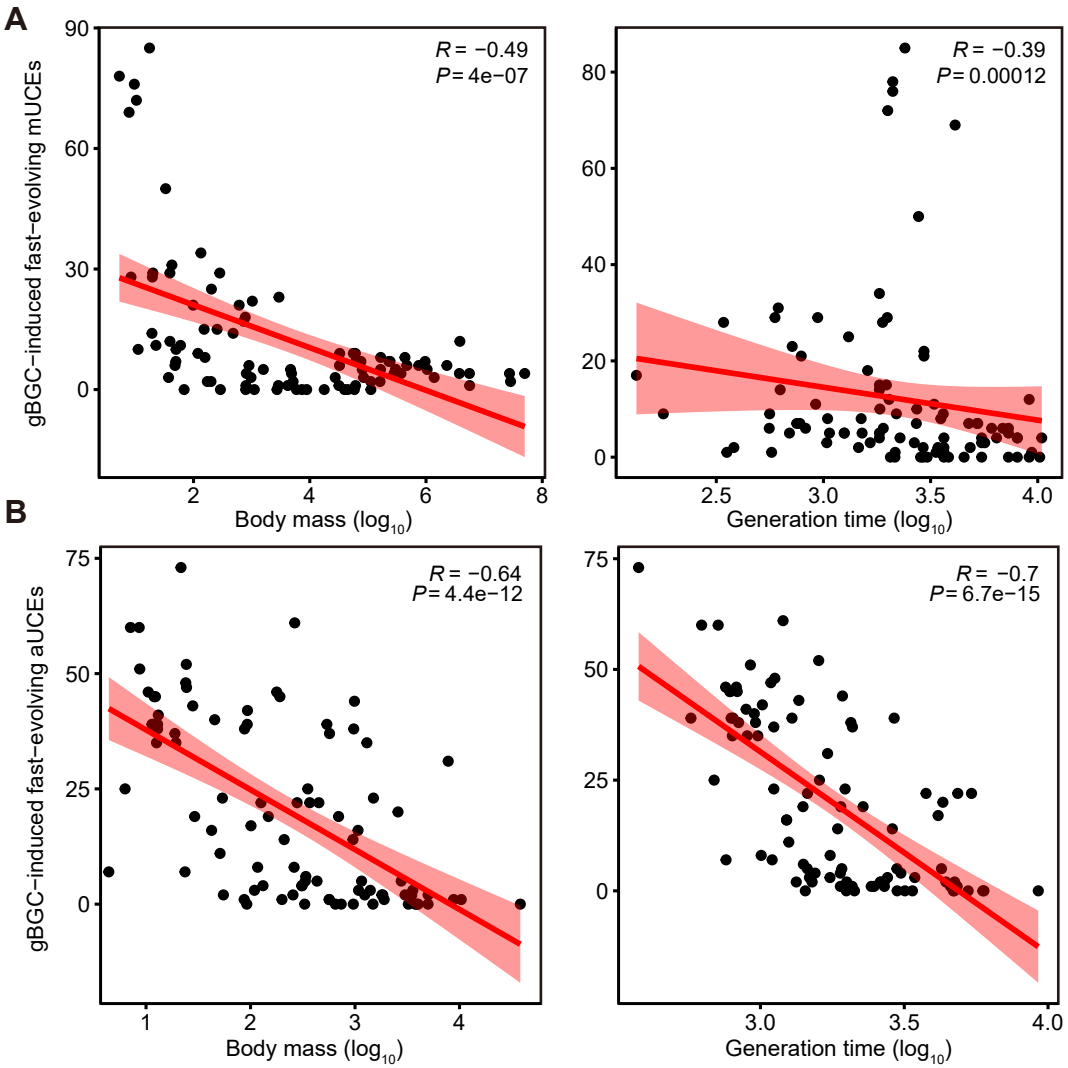
substitution numbers in fast-evolving UCEs of all species. (F) Distribution of substitution numbers in gBGC-induced fast-evolving UCEs of all species. The substitution numbers of each fast-evolving UCE and gBGC-induced fast-evolving UCE were counted according to the mutation direction, including S→W type and W→S type. The X axis indicates the number of W→S substitutions [n(W→S)]. S→W substitutions were scored as negative values.



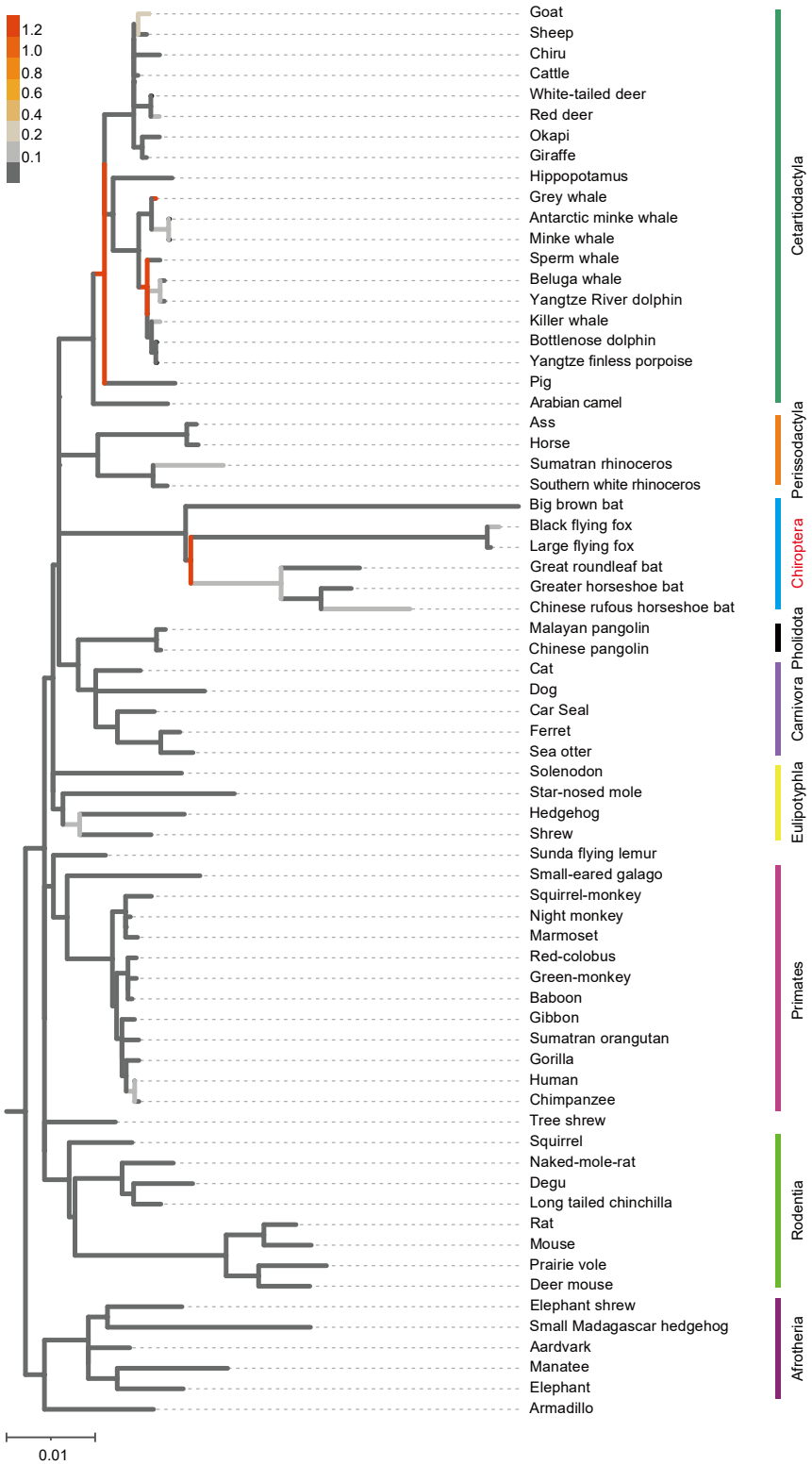

**Supplemental Fig. S10 Negative associations between fast-evolving UCEs and two life-history traits.** (A) Negative correlations between the number of fast-evolving mUCEs and body mass (left) and generation time (right) in mammals. (B) Negative correlations between the number of fast-evolving aUCEs and body mass (left) and generation time (right) in birds.



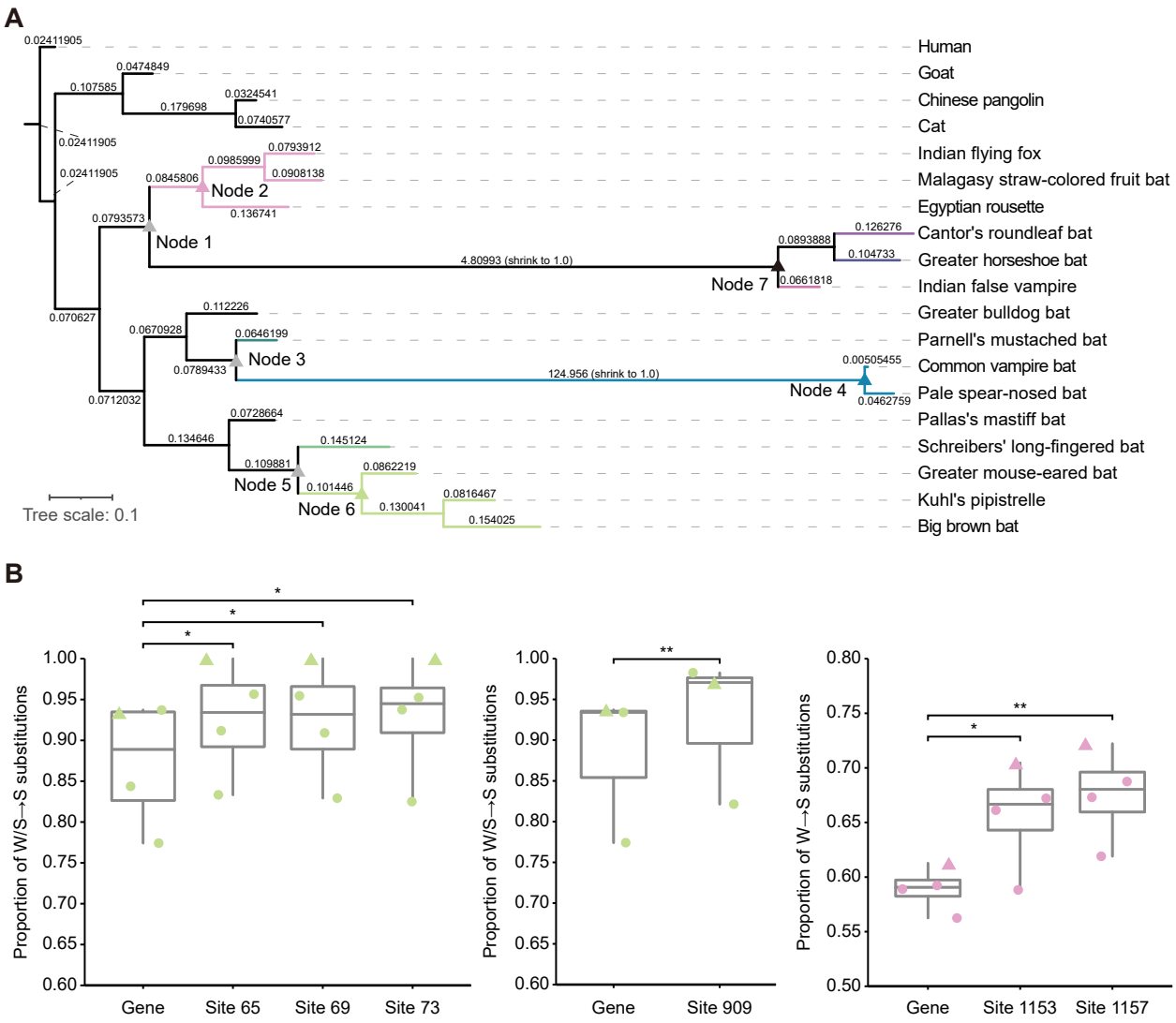

**Supplemental Fig. S11 Associations between other confounding factors and the number of fast-evolving UCEs.** (A) Positive correlation between the number of fast-evolving UCEs and genome-wide neutral substitution rate in mammals (left) and birds (right).(B) Correlation between the number of fast-evolving UCEs and assembly size of mammalian (left) and avian (right) species. (C) Correlation between the number of lineage-specific fast-evolving UCEs and the number of species included in mammalian (left) and avian (right) lineages.




**Supplemental Fig. S12 Negative associations between gBGC-induced fast-evolving UCEs and two life-history traits.** (A) Negative correlations between the number of gBGC-induced fast-evolving mUCEs and body mass (left) and generation time (right) in mammals. (B) Negative correlations between gBGC-induced fast-evolving aUCEs and body mass (left) and generation time (right) in birds.

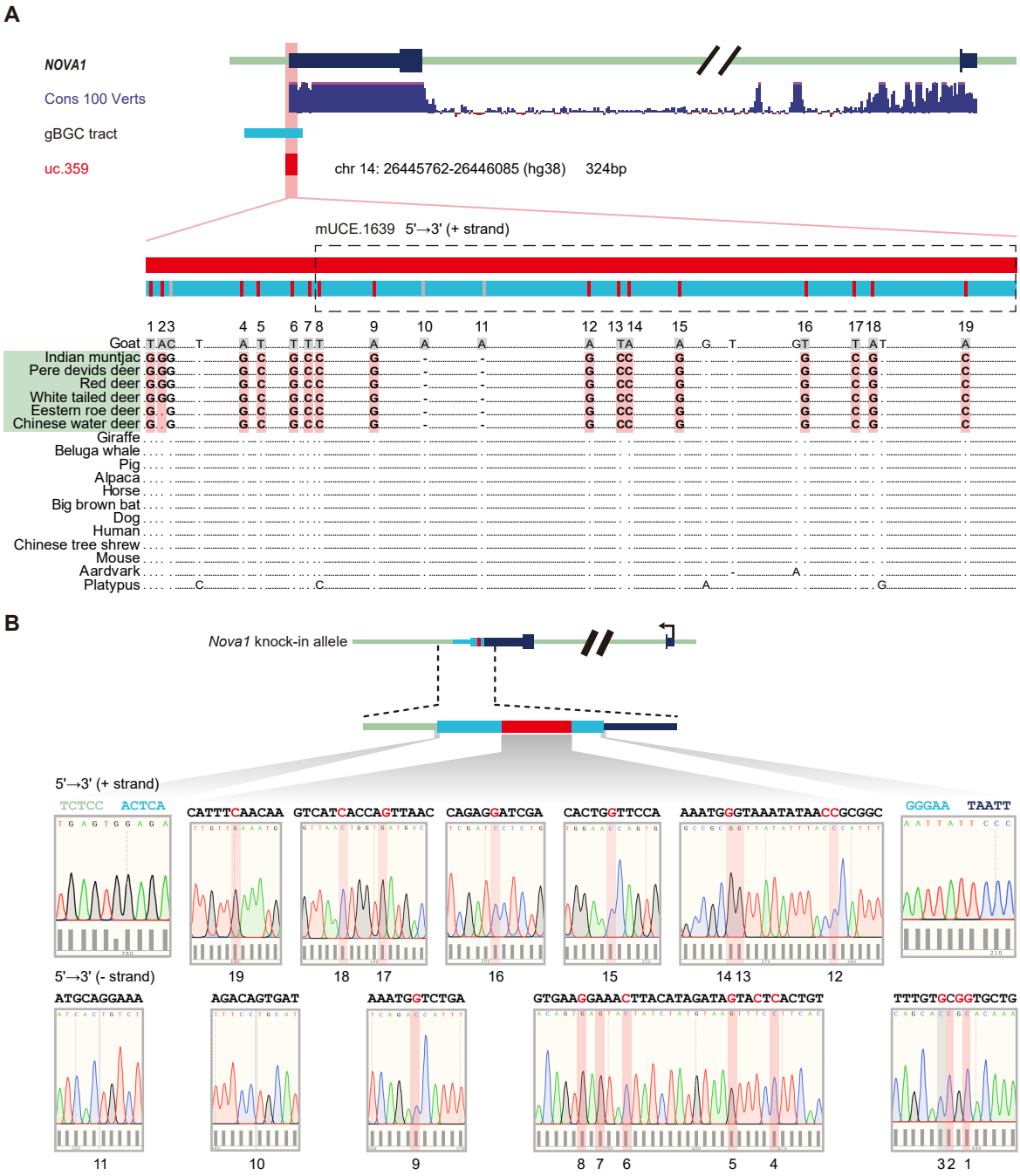



**Supplemental Fig. S13 Non-synonymous substitution rate ($d_N$) of *ZNF536* in 69 mammalian species.** The $d_N$ shows that *ZNF536* gene in bat species has accumulated many mutations. The branch lengths represent $d_N$ values, while the different $d_N/d_S$ ratios are shown in different colors.

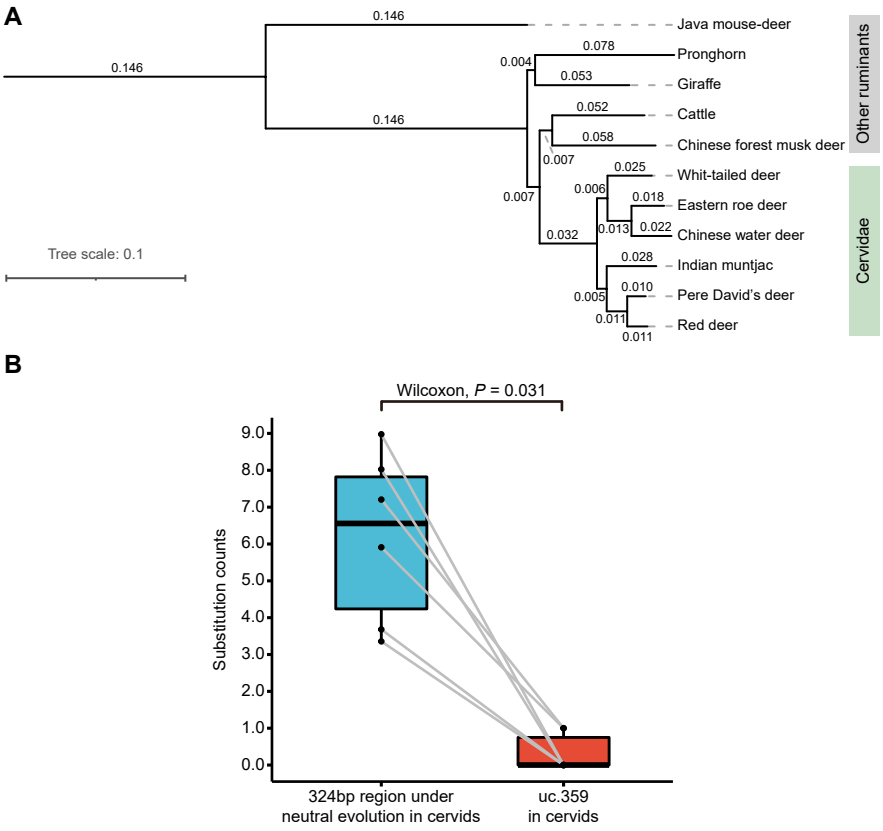



**Supplemental Fig. S14 Positively selected sites induced by gBGC in chiropteran *ZNF536*.**

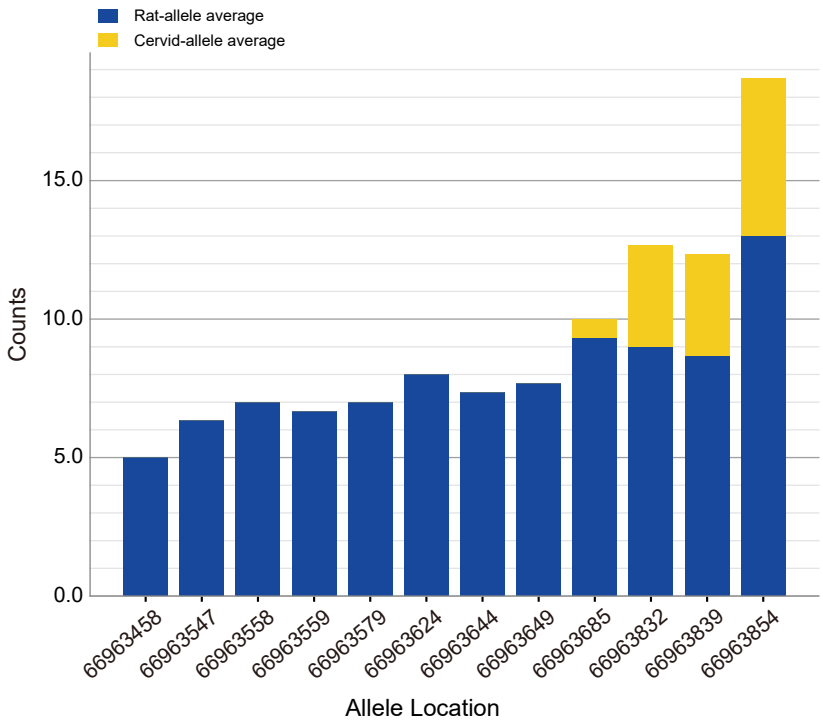
(A) Four lineages tested for positively selected sites within the *ZNF536* gene. Branch lengths represent $d_N/d_S$ ratios of the *ZNF536* gene. Test nodes are indicated by triangles in different colors (node2, 4, 6, 7), while the compared ancestral nodes are shown in gray triangles (node1, 3, 5). (B) Comparison of the proportion of W/S→S substitutions in the flanking 100 sites (300-bp) of six positively selected sites inferred by PAML and that in the entire *ZNF536* gene. Triangles present the corresponding reconstructed ancestral nodes of each lineage in (A). * $P <$ 0.1; ** $P < 0.01$.




**Supplemental Fig. S15 Cervid-specific substitutions in uc.359.** (A) Alignment of uc.359. Cervid-specific mutations are show in bold and the W→S mutations are further highlighted in red. (B) Validation of gene-edited heterozygous rats. Sanger sequencing confirmed the correct replacement of the cervid allele. Cervid-specific mutations are indicated by the corresponding number in (A). All W→S mutations are highlighted in red, while other cervid-specific mutations are highlighted in gray.




**Supplemental Fig. S16 Purifying selection analysis of cervid uc.359.** (A) Estimated neutral evolutionary rate of 11 ruminant species based on 4-fold degenerate sites. Branch length presents the number of substitutions per site. (B) Comparison of substitution counts between 324bp region under neutral evolution and uc.359 in six cervid species.




**Supplemental Fig. S17 Allele counts in Knock-in heterozygous rats.** Counts of rat-allele and cervid-allele on 12 fixed divergent sites (on Chr6, rn6) in heterozygous rats.

## Supplemental Tables

**Supplemental Table S1** Eutherian species and the correspomding data analyzed in this study.

**Supplemental Table S2** Avian species and the correspomding data analyzed in this study.

**Supplemental Table S3** Genomic coordinates of mammalian UCEs.

**Supplemental Table S4** Genomic coordinates of avian UCEs.

**Supplemental Table S5** Additional assemblies analyzed in this study.

**Supplemental Table S6** Human and chicken transcriptomes used in gene expression analyses.

**Supplemental Table S7** GO enrichment of mammalian UCEs.

**Supplemental Table S8** GO enrichment of avian UCEs.

**Supplemental Table S9** Counts and percentages of fast-evolving UCEs and gBGC-induced fast-evolving UCEs in each lineage.

**Supplemental Table S10** Predicted gBGC prevalence in UCEs by phastBias.

**Supplemental Table S11** 58 fast-evolving UCEs around *ZNF536* after interspecific synteny filtering.

**Supplemental Table S12** 19 assemblies used in *ZNF536* analysis.

**Supplemental Table S13** Positively selected sites inferred by CODEML in PAML.

**Supplemental Table S14** W/S→S substitutions within positively selected sites and their flanking regions and *ZNF536*.

**Supplemental Table S15** 811 bp target sequence in rats and the corresponding 813 bp homologous donor sequence in cervids.

**Supplemental Table S16** Genes with significantly altered expression in gene-edited rats.

**Supplemental Table S17** KEGG and GO enrichment of genes with significantly alterced expression in editing rats.

**Supplemental Table S18** 7 ancient gBGC-induced mUCEs.

**Supplemental Table S19** Predicted transcription factor binding motifs within the mUCE.1304 othologous sequences in humans and chikcens.

**Supplemental Table S20** The oligonucleotides used for sgRNA expression vectors.

**Supplemental Table S21** Primers used for amplifying and sequencing CRISPR/Cas9-induced knock-in segment in rats.

## Supplemental Data

**Supplemental Data S1** 19-way *ZNF536* coding sequence alignment.
The canonical conserved transcript of human *ZNF536* (ENST00000355537.4) were used as the reference to obtain alignment.

**Supplemental Data S2** Human and chicken orthologous sequences of mUCE.1304.

**Supplemental Data S3** $d_N/d_S$ value of *ZNF536* in 69 mammals.
Phylogenetic tree of 69 selected mammals in Newick format with number representing $d_N/d_S$ value of *ZNF536.*

**Supplemental Data S4** 19-way *ZNF536* coding sequence alignment used for $d_N/d_S$ analysis in PAML.

## References


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

Fu W, Wang R, Nanaei HA, Wang J, Hu D, Jiang Y. 2022. RGD v2.0: a major update of the ruminant functional and evolutionary genomics database. *Nucleic Acids Res* **50**: D1091-D1099.

Gonzalez JN, Zweig AS, Speir ML, Schmelter D, Rosenbloom KR, Raney BJ, Powell CC, Nassar LR, Maulding ND, Lee CM et al. 2021. The UCSC Genome Browser database: 2021 update. *Nucleic Acids Research* **49**: D1046-D1057.

Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12**: 656-664.

Kirilenko BM Munegowda C Osipova E Jebb D Sharma V Blumer M Morales AE Ahmed AW Kontopoulos DG Hilgers L et al. 2023. Integrating gene annotation with orthology inference at scale. *Science* **380**: eabn3107.

Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**: 268-274.