



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

Anguo Liu, Nini Wang, Guoxiang Xie, et al.

Genome Res. 2023 33: 1673-1689 originally published online October 26, 2023

Access the most recent version at doi:[10.1101/gr.277784.123](https://doi.org/10.1101/gr.277784.123)

References

This article cites 104 articles, 25 of which can be accessed free at:
<http://genome.cshlp.org/content/33/10/1673.full.html#ref-list-1>

Creative Commons License

This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see <https://genome.cshlp.org/site/misc/terms.xhtml>). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

Email Alerting Service

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).

Comprehensive immune receptor profiling.
Discover the **DriverMap™ AIR Assay** difference.



To subscribe to *Genome Research* go to:
<https://genome.cshlp.org/subscriptions>

Research

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

Anguo Liu,^{1,2,7} Nini Wang,^{1,3,7} Guoxiang Xie,^{1,2,7} Yang Li,^{1,2} Xixi Yan,^{1,2} Xinmei Li,^{1,2} Zhenliang Zhu,^{2,4,5} Zhuohui Li,^{1,2} Jing Yang,^{2,4,5} Fanxin Meng,¹ Mingle Dou,^{1,2} Weihuang Chen,¹ Nange Ma,^{1,2} Yu Jiang,^{1,2,6} Yuanpeng Gao,^{2,4,5} and Yu Wang^{1,2}

¹Key Laboratory of Animal Genetics, Breeding and Reproduction of Shaanxi Province, College of Animal Science and Technology, Northwest A&F University, Yangling, Shaanxi 712100, China; ²Key Laboratory of Livestock Biology, Northwest A&F University, Yangling, Shaanxi 712100, China; ³Faculty of Mathematics and Natural Sciences, University of Cologne, and Cologne Excellence Cluster for Cellular Stress Responses in Aging-Associated Diseases (CECAD), University Hospital Cologne, Cologne 50931, Germany; ⁴College of Veterinary Medicine, Northwest A&F University, Yangling, Shaanxi 712100, China; ⁵Key Laboratory of Animal Biotechnology, Ministry of Agriculture, Northwest A&F University, Yangling, Shaanxi 712100, China; ⁶Center for Functional Genomics, Institute of Future Agriculture, Northwest A&F University, Yangling, Shaanxi 712100, China

Ultraconserved elements (UCEs) are the most conserved regions among the genomes of evolutionarily distant species and are thought to play critical biological functions. However, some UCEs rapidly evolved in specific lineages, and whether they contributed to adaptive evolution is still controversial. Here, using an increased number of sequenced genomes with high taxonomic coverage, we identified 2191 mammalian UCEs and 5938 avian UCEs from 95 mammal and 94 bird genomes, respectively. Our results show that these UCEs are functionally constrained and that their adjacent genes are prone to widespread expression with low expression diversity across tissues. Functional enrichment of mammalian and avian UCEs shows different trends indicating that UCEs may contribute to adaptive evolution of taxa. Focusing on lineage-specific accelerated evolution, we discover that the proportion of fast-evolving UCEs in nine mammalian and 10 avian test lineages range from 0.19% to 13.2%. Notably, up to 62.1% of fast-evolving UCEs in test lineages are much more likely to result from GC-biased gene conversion (gBGC). A single cervid-specific gBGC region embracing the uc.359 allele significantly alters the expression of *Noval* and other neural-related genes in the rat brain. Combined with the altered regulatory activity of ancient gBGC-induced fast-evolving UCEs in eutherians, our results provide evidence that synergy between gBGC and selection shaped lineage-specific substitution patterns, even in the most constrained regulatory elements. In summary, our results show that gBGC played an important role in facilitating lineage-specific accelerated evolution of UCEs, and further support the idea that a combination of multiple evolutionary forces shapes adaptive evolution.

[Supplemental material is available for this article.]

Understanding the genetic basis of species diversity and their environmental adaptations is a long-standing issue in evolution. Sequence conservation is a powerful signature for identifying functional elements, and base component changes in functional elements may contribute to species' adaptive evolution (Boffelli et al. 2004; Kvon et al. 2016; Sackton et al. 2019; Uebbing et al. 2021). One such extreme example is a set of 481 ultraconserved elements (UCEs), which were originally defined as sequences that are >200 bp and absolutely conserved across mouse, rat, and human genomes (Bejerano et al. 2004). These UCEs were reported to maintain strong conservation in other lineages, even beyond mammals (Bejerano et al. 2004; de la Calle-Mustienes et al. 2005; Miller et al. 2007). Additionally, single-nucleotide polymorphisms (SNPs) are rare in UCEs, which further indicates that these ele-

ments are under extreme negative selection (Drake et al. 2006; Chen et al. 2007; Katzman et al. 2007).

UCEs have also been found to encompass various functions, including enhancer, promoter, splicing, and repressive activities (Pennacchio et al. 2006; Poitras et al. 2010; Snetkova et al. 2022). In vivo transgenic analysis showed that approximately half of non-coding UCEs serve as tissue-specific enhancers that regulate gene expression during development (Visel et al. 2008). Functional experiments subsequently proved that UCEs are important for organisms, even though alterations within some UCEs have subtle or nonvisible effects (Braconi et al. 2011; Vannini et al. 2017; Dickel et al. 2018; Snetkova et al. 2021). However, many extremely conserved regions are still altered by some lineage-specific mutations that lead to phenotypic changes (Booker et al. 2016; Holloway et al. 2016; Feigin et al. 2019). Human highly accelerated

⁷These authors contributed equally to this work.

Corresponding authors: wang_yu@nwfau.edu.cn, gaoyuanpeng1990@163.com

Article published online before print. Article, supplemental material, and publication date are at <https://www.genome.org/cgi/doi/10.1101/gr.277784.123>.

© 2023 Liu et al. This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see <https://genome.cshlp.org/site/misc/terms.xhtml>). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

region 1 (HAR1) is a remarkable example in which the sequence is extremely conserved among nonhuman vertebrates but rapidly evolved in humans. Functional analysis indicated that HAR1 is involved in the unique neurodevelopment of humans (Pollard et al. 2006b). Similarly, human accelerated conserved noncoding sequence 1 (HACNS1) is another example; it is thought to serve as a human-specific enhancer that contributed to the evolution of human limb features (Prabhakar et al. 2008; Dutrow et al. 2022).

Recently, researchers also found that several UCEs have changed in some mammalian lineages despite their extreme conservation across other mammals (Wang et al. 2019b; Hecker and Hiller 2020; Jebb et al. 2020). These studies highlighted the important role of accelerated evolution of highly conserved regions. Therefore, it may be possible to detect fast-evolving UCEs with specific functional alterations in mammals and birds that potentially contributed to mammalian and avian traits.

Although interest in lineage-specific accelerated regions has mainly been attributed to their potential relevance to specific functional alterations and adaptive evolution, GC-biased gene conversion (gBGC) is an alternative explanation that could also increase local fixation rate and thus confound detection of positive selection (Galtier and Duret 2007; Gotea and Elnitski 2014). gBGC is a recombination-associated process that favors the fixation of G/C (strong [S]) alleles over A/T (weak [W]) alleles, thus leading to elevated GC content in affected regions (Duret and Galtier 2009a). To date, gBGC has been found to be widespread in mammals and birds and contributes to GC content variation, rapid evolution of sequences, and even fixation of deleterious mutations (Galtier et al. 2009; Glémin 2010; Pessia et al. 2012; Capra et al. 2013; Lartillot 2013; Lachance and Tishkoff 2014; Weber et al. 2014; Borges et al. 2019). Some studies have also found that a combination of gBGC and directional selection contributed to the accelerated evolution of fast-evolving elements; therefore, gBGC could drive the fixation of mutations that potentially affect species' adaptation (Pollard et al. 2006a,b; Prabhakar et al. 2008; Kostka et al. 2012; Ferris et al. 2018; Feigin et al. 2019; Mangan et al. 2022). Notably, substitutions of the above two human accelerated sequences showed clear A/T $\rightarrow$ G/C (W $\rightarrow$ S) bias. All 18 human-specific mutations of the 118-bp HAR1 region were W $\rightarrow$ S conversions (Pollard et al. 2006b). Similarly, the 81-bp HACNS1 subregion has 13 human-specific substitutions, 12 of which were W $\rightarrow$ S conversions (Prabhakar et al. 2008). The mutations in both regions were initially inferred to result from gBGC and directional selection in human evolution, whereas subsequent analysis and functional experiments supported that the evolution of HACNS1 in the human lineage results from a loss of function driven by gBGC (Duret and Galtier 2009b; Sumiyama and Saitou 2011). Although a previous study delineated the prevalence of gBGC along human and chimpanzee genomes (Capra et al. 2013), most studies mainly focused on protein-coding genes or the general impact of gBGC in humans and a few other lineages of interest (Kostka et al. 2012; Mugal et al. 2013; Figueat et al. 2014; Rousselle et al. 2019). However, the functional consequences of identified acceleration regions that can be purely explained by gBGC remain unknown.

Here, benefitting from the increased availability of sequenced genomes, we comprehensively identified reliable sets of UCEs for both mammals and birds. Focused on lineage-specific fast-evolving UCEs, we investigated the potential driving role of gBGC among them and further explored the functional consequences of gBGC-induced fast-evolving UCEs. Our study provided evidence that gBGC-induced acceleration of UCEs may contribute

to lineage-specific alterations, and highlighted that a combination of multiple evolutionary forces facilitates regulatory alterations that affect adaptive evolution.

Results

Identification of UCEs in mammals and birds

Traditional UCE detection mainly relies on a strict but arbitrary filter threshold. Such a direct filtering process often overlooked the issue of inconsistent evolutionary rate among different species. As more genomes are included in alignment, it becomes more challenging to achieve 100% identity across all (or most) genomes. In other words, the only way to find enough UCEs with high confidence is by reducing either (1) the length threshold or (2) the number of genomes that must maintain 100% identity in candidate regions (Christmas et al. 2023).

Here, to identify reliable mammalian UCEs (mUCEs) and avian UCEs (aUCEs) using numerous species, we developed a comparative genomics approach that considers potential species-specific mutations and assembly errors without compromising length or quantity (Fig. 1). First, we constructed comprehensive multiple genome alignments for 95 eutherian and 94 avian species (Supplemental Fig. S1; Supplemental Tables S1, S2), which covered all orders of the two taxa. We then reconstructed mammalian and avian ancestral genomes by selecting 10 high-quality genomes from different orders of Mammalia and Aves, respectively (Fig. 1A). Given the preconceived notion that UCEs are expected to be extremely conserved in most lineages and thus in most of 10 genomes used in ancestral reconstruction, each reconstructed ancestral base in potential UCE regions is expected to be consistent and independent of the given phylogeny. Next, all species' genomes were simulated as short reads and aligned to the corresponding ancestral genome to obtain absolutely conserved sequences. The sequences that were completely aligned to the ancestral genome without any mismatches were extracted from the ancestral genome and merged to obtain as long of a candidate UCE as possible (≥ 100 bp at least) (Fig. 1B). This step would further reduce the possibility of false ultraconserved regions in ancestral state reconstruction. In total, 63,678 candidate mUCEs and 147,348 candidate aUCEs were obtained. In our candidate mUCE set, 474 of the original UCEs (Bejerano et al. 2004) were also included (98.5%, 474/481) (see Supplemental Methods and Results), which indicated the reliability of our candidates for further filtering (Supplemental Fig. S2).

Next, to identify a set of precise UCEs, we divided 95 mammals and 94 birds into eight and three branches, respectively, based on mammalian and avian phylogeny (see Methods) to ensure that each branch contained at least three species (Fig. 1A). Using more stringent conditions extended from the original definition of UCEs (Bejerano et al. 2004), the candidate UCEs that were >200 bp and 100% identical within at least one branch (three or more species) with at least 80% identity across all branches were defined as final UCEs (Fig. 1B). In total, 2191 mUCEs and 5938 aUCEs were finally identified (Supplemental Tables S3, S4). The longest mUCE and aUCE were 806 bp and 1773 bp, respectively, and the median mUCE and aUCE were 250 bp and 254 bp, respectively. Notably, our mUCEs are largely independent of the 481 original UCEs; only 194/481 original UCEs were included among our final mUCEs (Supplemental Figs. S2, S3). One representative example is mUCE.1968, which is adjacent to the previously identified uc.428 (mUCE.1969). This mUCE showed ultraconservation in mammals despite several lineage-specific mutations in rodents

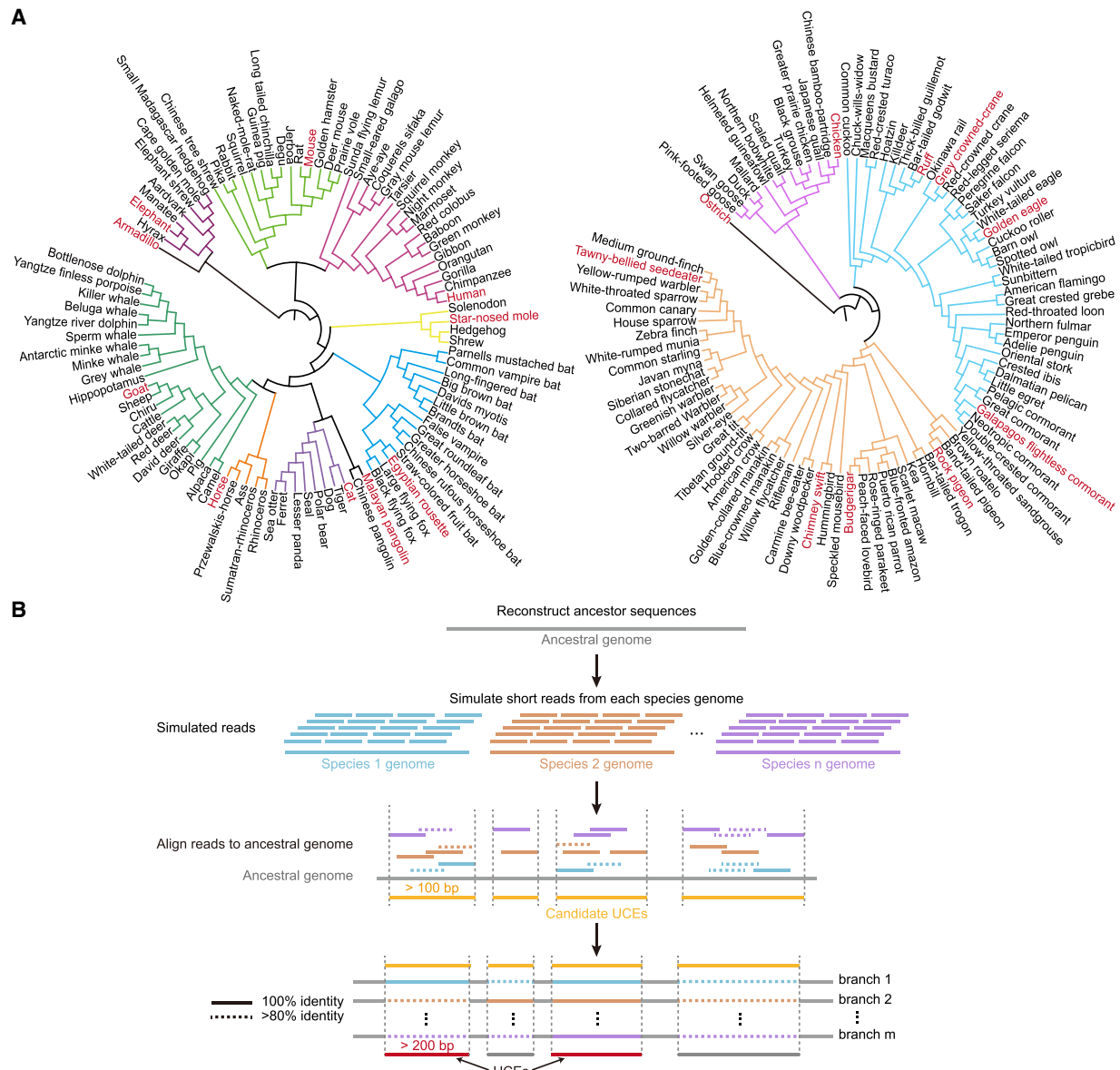


Figure 1. UCE identification flowchart. (A) Phylogenetic trees of 95 mammals (*left*) and 94 birds (*right*). The trees were constructed by IQ-TREE (Nguyen et al. 2015) using fourfold degenerate sites and the maximum likelihood method. Different branches are shown in different colors, and the names of the species whose genomes were used to reconstruct ancestral genomes are shown in red. (B) Pipeline for identifying mammalian UCEs (mUCEs) and avian UCEs (aUCEs).

and pangolins (Supplemental Fig. S4). Together, these results showed the effectiveness of our identification pipeline and indicated that additional true UCEs could only be captured when they were aligned using numerous genomes.

mUCE and aUCE characteristics

To further verify the credibility of our UCEs and their potential functional importance, we conducted a series of investigations on their characteristics. Consistent with previous studies (Bejerano et al. 2004; Zhang et al. 2014; Seki et al. 2017), the genomic distribution of our mUCEs and aUCEs showed that more UCEs (73.5% and 76.1%) were located in noncoding regions (Fig. 2A). When we further investigated our mUCEs, more than half (54.6%, 1196/

2191) were 100% identical within at least two branches, and 48 mUCEs were absolutely conserved among all branches (Fig. 2B).

To further explore the changes of ultraconservation along vertebrate evolution, we examined the conservation of our UCEs in all vertebrate taxa, including representatives of Reptilia, Amphibia, Osteichthyes, Chondrichthyes, and Cyclostomata (list in Supplemental Table S5). Most (>80%) mUCEs and aUCEs had homologous sequences in Reptilia, but the number sharply decreased with more distant lineages. Approximately half of UCEs were conserved in Amphibia and Coelacanthimorpha and partially maintained in Chondrichthyes (~30%), and there was nearly nonexistent homology in Cyclostomata (jawless vertebrates) (Fig. 2C; Supplemental Fig. S5A). These results based on conservation analysis further supported the idea that many UCEs were

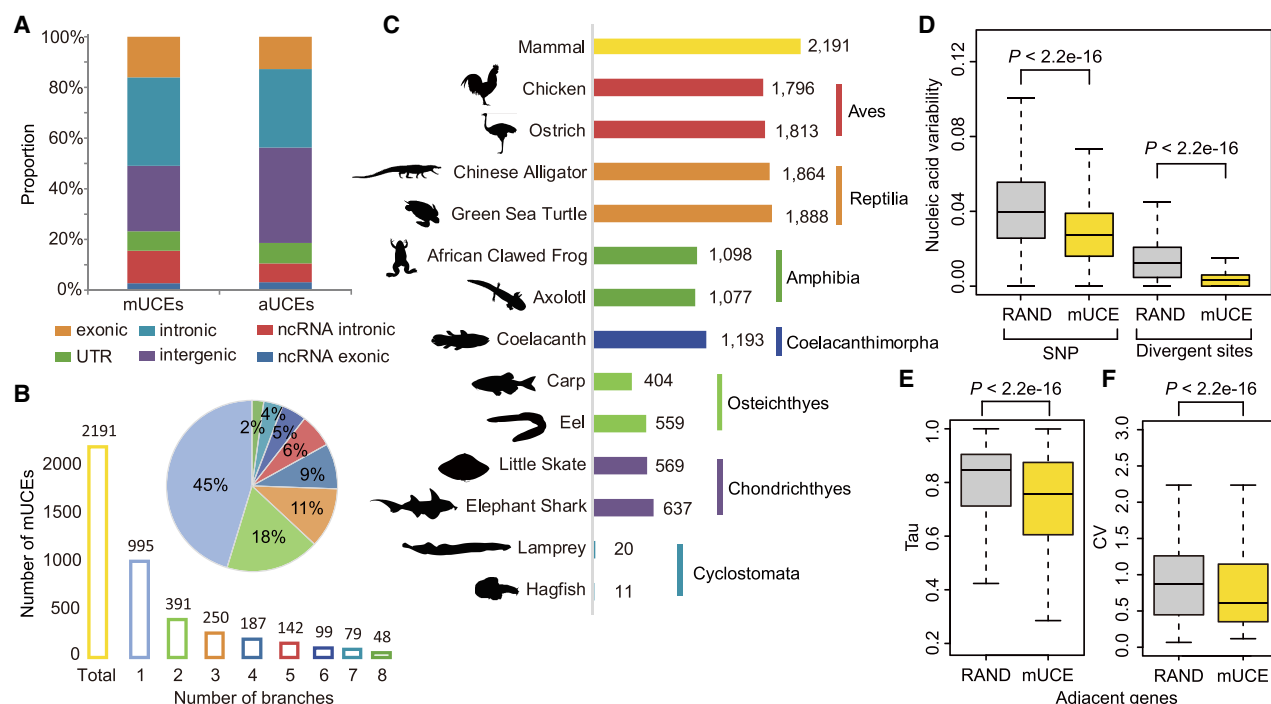


Figure 2. UCE characteristics. (A) Genomic distribution of mUCEs and aUCEs. (B) Number of UCEs that were 100% identical to the ancestor when different numbers of mammalian branches were considered. The pie chart shows the proportion of UCEs when different branches were included. (C) Numbers of homologous sequences of mUCEs present in different vertebrate taxa. (D) Enrichment of human SNPs and counts of divergent sites between humans and chimpanzees in UCEs and randomly selected genomic regions (RANDs). (E) Tissue-specific expression index (τ) of UCE-adjacent genes and RAND-adjacent genes. (F) Gene expression diversity value (CV) of UCE-adjacent genes and RAND-adjacent genes.

potential innovations produced in different evolutionary stages after divergence from the common ancestor with fish and finally became ultraconserved in birds and mammals (Bejerano et al. 2004; Stephen et al. 2008). However, further experimental validations are necessary to investigate their regulatory roles and elucidate these evolutionary scenarios.

To test whether UCEs resulted from selection or a low mutation rate, we then detected the human SNPs and divergent sites between humans and chimpanzees in the mUCEs and randomly selected genomic regions (RANDs). For aUCEs, we performed the same analyses using chicken SNPs and divergent sites of chickens and turkeys. In both taxa, the SNP enrichment showed a significant difference (P -value $< 2.2 \times 10^{-16}$, Wilcoxon rank-sum test) between UCEs and RANDs, irrespective of the SNP frequency (Fig. 2D; Supplemental Figs. S5B, S6A,B). Moreover, the counts of divergent sites were also significantly depleted in UCE regions (P -value $< 2.2 \times 10^{-16}$, Wilcoxon rank-sum test) (Fig. 2D; Supplemental Fig. S5B). This slow evolutionary pace of low SNP density and low level of interspecific divergence is consistent with purifying selection, but it might also result from low mutation rate. By analyzing site frequency spectra, we further found that SNPs within UCEs segregate at a lower allelic frequency than SNPs within RANDs, which shows that UCEs are under purifying selection (Supplemental Fig. S6C, D). These results indicated the credibility of our UCEs because of the expectation that UCEs do not result from low-mutation regions of the genome but are subject to extremely strong purifying selection during evolution (Drake et al. 2006; Katzman et al. 2007).

Because UCEs are often situated in the vicinity of functionally important genes and potentially exert *cis*-regulatory effects on their expression (Snetkova et al. 2022), we explored the expression

patterns of the genes adjacent to our UCEs. Based on 55 transcriptomes covering 11 tissues from humans and 62 transcriptomes covering 17 tissues from chickens (Supplemental Table S6), we calculated the gene expression tissue-specific index (τ) and gene expression diversity index (CV). Both the τ and CV values of UCE-adjacent genes were significantly lower (P -value $< 2.2 \times 10^{-16}$, Wilcoxon rank-sum test) than those of RAND-adjacent genes; this indicated that the UCE-adjacent genes were more likely to have widespread expression and possess lower expression diversity across tissues (Fig. 2E,F; Supplemental Fig. S5C,D). These results showed that UCEs are preferentially located near or within genes that play important biological functions. Gene Ontology (GO) analysis of UCE-adjacent genes further confirmed that conclusion. All UCEs were commonly enriched near genes that play multiple roles in the regulation of DNA transcription and RNA synthesis. Additionally, genes near UCEs were enriched for terms associated with regulating the development of the nervous system in mammals, especially neuron differentiation and regulation. In birds, they were more enriched for terms associated with body development, including embryonic development, bone development, and development of other organs (Supplemental Fig. S7; Supplemental Tables S7, S8). Upon comparing the simplified GO terms of two sets, we determined that the final similarity index was 0.672, indicating a moderate difference between two GO enrichment sets. Compared with previous independent enrichment analyses of human–rodent UCEs and avian-specific highly conserved elements (Bejerano et al. 2004; Seki et al. 2017), our direct comparison of mUCEs and aUCEs at the same scale further highlighted a moderate difference in enrichment trends of regulatory elements between the two taxa; this reflects the potential

importance of UCEs in the evolution of lineage adaptations. Collectively, all the above features of our UCEs indicated their high credibility and potential importance, which should be further explored.

Accelerated evolution of UCEs in specific mammal and bird lineages

Significantly low variability indicates that natural variations within UCEs may potentially affect their functions (Drake et al. 2006). Previous studies found that some of the original 481 UCEs rapidly evolved in different lineages (Wang et al. 2019b; Hecker and Hiller 2020; Jebb et al. 2020). In our study, only 48 detected mUCEs were identical in all eight mammalian branches, which further indicated that the majority of UCEs were relatively variable.

To evaluate UCE variability and identify those with the most potential to have functional impacts, we detected the fast-evolving UCEs in selected mammalian and avian lineages using CONACC mode and the LRT method (Hubisz et al. 2011). To minimize the effects of incongruence in the phylogenetic placement of several neoavian branches, we carefully selected 10 avian lineages and tested lineage-specific fast-evolving UCEs based on our avian phylogeny and a reference phylogeny (Jarvis et al. 2014), respectively. Notably, the results from these two different avian phylogenetic trees were highly consistent (Supplemental Fig. S9A–D), indicating the relative robustness of lineage-specific fast-evolving UCEs in 10 test avian lineages. The final results showed that 0.59%–13.24% and 0.19%–6.33% of UCEs underwent accelerated evolution in

mammal and bird test lineages, respectively (Fig. 3A; Supplemental Table S9). We also found negative associations between the number of fast-evolving UCEs and two life-history traits, body mass and generation time (Supplemental Fig. S10), although there are other confounding factors (such as mutation rate) that could potentially contribute to the number of fast-evolving UCEs (Supplemental Fig. S11).

Driving role of gBGC in shaping fast-evolving UCEs

UCE conservation provides the opportunity to explore underlying mutational mechanisms of lineage-specific accelerated evolution. We next considered gBGC, which can mimic selection-driven accelerated evolution in specific lineages by fixing GC-biased mutations in evolutionarily conserved regions (Galtier and Duret 2007; Kostka et al. 2012). We first assessed whether there was a GC-biased mutation trend among UCEs in different lineages. Using traditional summary statistics, we counted the number of substitutions in all mUCEs in nine mammalian lineages according to two mutation directions, $S \rightarrow W$ ($G/C \rightarrow A/T$) and $W \rightarrow S$ ($A/T \rightarrow G/C$), and excluded potential noise from other mutation directions. In all lineages, the substitutions in the two mutation directions were almost equal in all UCEs, but substitutions in fast-evolving UCEs were clearly biased toward $W \rightarrow S$ (P -value $< 2.2 \times 10^{-16}$, Wilcoxon rank-sum test) (Fig. 3B,C; Supplemental Figs. S8, S9E,F). Such results supported our hypothesis that gBGC potentially influenced the lineage-specific accelerated evolution of UCEs.

To further determine the prevalence of gBGC in shaping fast-evolving UCEs, we used phastBias (Capra et al. 2013) to predict

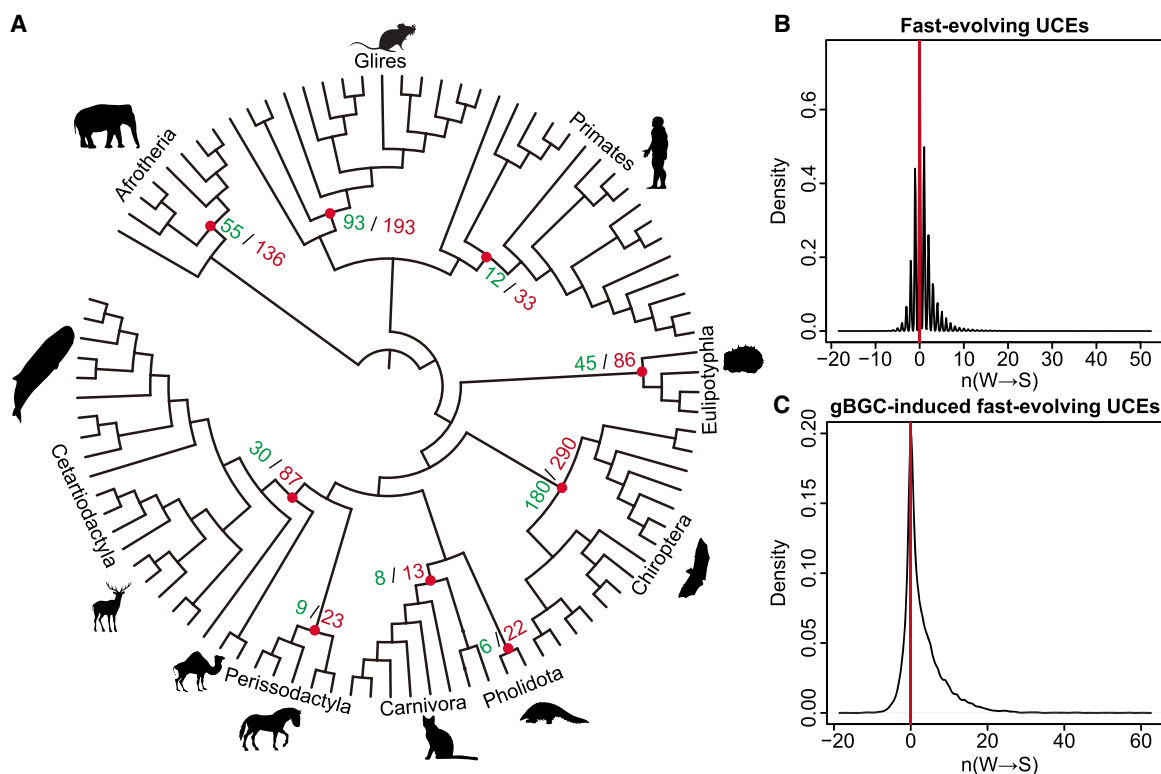


Figure 3. Accelerated evolution and gBGC of UCEs in mammals. (A) Number of fast-evolving UCEs (red) and gBGC-induced fast-evolving UCEs (green) in specific mammalian lineages (red dots). (B) Distribution of substitution numbers in fast-evolving UCEs of all species. (C) Distribution of substitution numbers 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 $S \rightarrow W$ type and $W \rightarrow S$ type. The x -axis indicates the number of $W \rightarrow S$ substitutions: $[n(W \rightarrow S)]$. $S \rightarrow W$ substitutions were scored as negative values.

informative regions of gBGC for all test lineages. Using default parameters, nearly all fast-evolving UCEs were informative gBGC regions (defined as regions where >50% of the sites were informative for gBGC) in all test lineages, but only a minority (10.9% on average) were located in gBGC tracts (defined as regions with posterior probability of being in a gBGC state >0.5) (Supplemental Table S10). This is explainable as only extremely strong or long-lasting gBGC events could be captured by this phylogenetic method (Glémin et al. 2015), which selectively omits many real gBGC events that do not fit the parameters because of weaker signals. Because clusters of biased substitutions potentially provide more direct evidence of a local influence from gBGC (Capra et al. 2013), to accurately identify gBGC-induced acceleration among UCEs, we used strict conditions for subsequent screening: Only those fast-evolving UCEs whose W → S substitutions were significantly greater (FDR < 0.05; Storer–Kim test) than S → W substitutions were identified as gBGC-induced fast-evolving UCEs (see Methods). Final statistics showed that, on average, 44.6% and 25.0% lineage-specific fast-evolving UCEs were much more likely to result from gBGC in mammals and birds, respectively, but the proportion of gBGC-induced fast-evolving UCEs in different lineages varied greatly (Fig. 3A; Supplemental Fig. S9A; Supplemental Table S9). Although we discarded questionable fast-evolving UCEs (P -value < 0.05 but FDR > 0.05), our results still showed that up to 62.1% of the fast-evolving UCEs were potentially affected by gBGC in test lineages (Supplemental Table S9). Furthermore, we also found that those lineages and species with a larger population size, smaller body mass, and shorter generation time had relatively larger proportions of gBGC-induced fast-evolving UCEs (Supplemental Fig. S12). This is reasonable as the strength of gBGC depends on its population-scaled coefficient ($B = 4N_e b$, where b is the gBGC coefficient) (Duret and Galtier 2009a). Given that effective population size (N_e) has been proposed to be negatively correlated with body mass and generation time, gBGC is expected to be more prevalent in small short-lived organisms (Romiguier et al. 2010; Webster and Hurst 2012; Lartillot 2013; Weber et al. 2014). Together, these results revealed that gBGC is the potential force responsible for the accelerated evolution of UCEs, and the magnitude of this effect is directly manipulated by different levels of gBGC.

gBGC promotes parallel alterations in genes and the proximal conserved elements

Strong gBGC can lead to severe W → S substitutions in both coding and flanking noncoding regions (Ratnakumar et al. 2010). Therefore, a conspicuous GC-biased substitution pattern in proximal conserved elements may be a potential proxy of the nearby effected genes. An illustrative example was provided by the zinc finger protein 536 (*ZNF536*) of Chiroptera, which had the highest amount of fast-evolving mUCEs (21/58) (Supplemental Table S11) within 1-Mb regions surrounding *ZNF536* after interspecific synteny filtering (see Methods), and 15/21 were considered to be gBGC-induced (Fig. 4A). Among them, one 432-bp representative UCE, mUCE.2036, is located in the *ZNF536* promoter region and represents a severe GC-biased landscape in Vespertilionid bats and other bat species. Within mUCE.2036, 42/45 Vespertilionidae-specific mutations were W → S substitutions that elevated GC content from 37.3% to 56.7% (Fig. 4B,C). In parallel, *ZNF536* has been subject to a massive accumulation of nonsynonymous substitutions in all branches of the Chiroptera clade, despite strong purifying selection (evidenced by the low ratio of nonsynonymous to synonymous substitution rates [d_N/d_S]) (Supplemental Fig. S13).

To further explore the substitution pattern in chiropteran *ZNF536*, we generated a 19-way *ZNF536* coding sequence alignment (Fig. 4D; Supplemental Data S1) including 15 bat species (listed in Supplemental Table S12) and found significantly elevated GC content (P -value = 0.0043, Mann–Whitney U -test) in bats compared with the outgroup species and the reconstructed common ancestors (Fig. 4E). Focused on the six bat species used in UCE identification, we further found that the proportion of W → S substitutions in the first and second codon positions (GC12) are significantly higher (P -value < 0.05, Mann–Whitney U -test) than those in the third codon positions (GC3) and in all codon positions (Fig. 4F), which is consistent with the gBGC-induced substitution pattern (Ratnakumar et al. 2010).

More W → S substitutions fixed in GC12 means more chances confounding detection of sites under positive selection. Hence, we further explored whether gBGCs have participated in facilitating this situation in chiropteran lineages. Among all amino acid sites inferred by PAML, we found that sites determined within a single lineage tend to cluster together (Supplemental Table S13). Additionally, six out of seven sites with high confidence (posterior probability > 0.95) possessed W → S substitutions in GC12, and their flanking 100 sites showed elevated W/S → S substitutions that provided direct evidence to be potentially affected by gBGC (Supplemental Fig. S14; Supplemental Table S14). Considering all of the described evidence, it is highly possible that gBGC played an important role in shaping a similar substitution pattern in both *ZNF536* and the proximal conserved elements.

One gBGC-induced cervid-specific fast-evolving UCE shows functional alteration

To further investigate whether there were functionalized fast-evolving UCEs induced by gBGC, we focused on a 271-bp UCE, mUCE.1639. This UCE is located in the 3' UTR of NOVA alternative splicing regulator 1 (*NOVA1*) and showed significant acceleration in cervids harboring 12 cervid-specific substitutions; 10/12 were W → S substitutions (Fig. 5A; Supplemental Fig. S15). mUCE.1639 includes the majority of uc.359 (271 bp/324 bp, 83.6%) (Bejerano et al. 2004), which was previously reported to be potentially involved in cancer-like growth of antlers (Wang et al. 2019b). When we examined uc.359, seven additional mutations were included, which were all W → S substitutions within a 54-bp narrow region except for one C → G substitution (Fig. 5A; Supplemental Fig. S15). Using phastBias (Capra et al. 2013), we further found that uc.359 is located in the sole gBGC tract in the 10-kb region surrounding uc.359. Notably, this gBGC tract was only captured in the common ancestor of cervids, although informative gBGC signal of this tract was detected in all cervid species. Further comparison of the substitution counts in uc.359 in Cervidae with those expected under neutral evolution indicated that uc.359 is still subject to strong purifying selection (Supplemental Fig. S16). Therefore, we speculated that a recombination event occurred in the cervid ancestor that shaped this cervid-specific GC-biased substitution pattern in uc.359 that potentially provides some kind of influence on fitness and suffers from selection pressure that resulted in perfect maintenance of this ancient gBGC event after lineage differentiation.

NOVA1 is a splicing regulator gene for which expression levels are tightly regulated. Its dysregulation has been considered to play a critical role in many different neurological diseases and cancers (Jensen et al. 2000; Ule et al. 2005; Xin et al. 2017). A recent study revealed that a single amino acid change in human *NOVA1* could

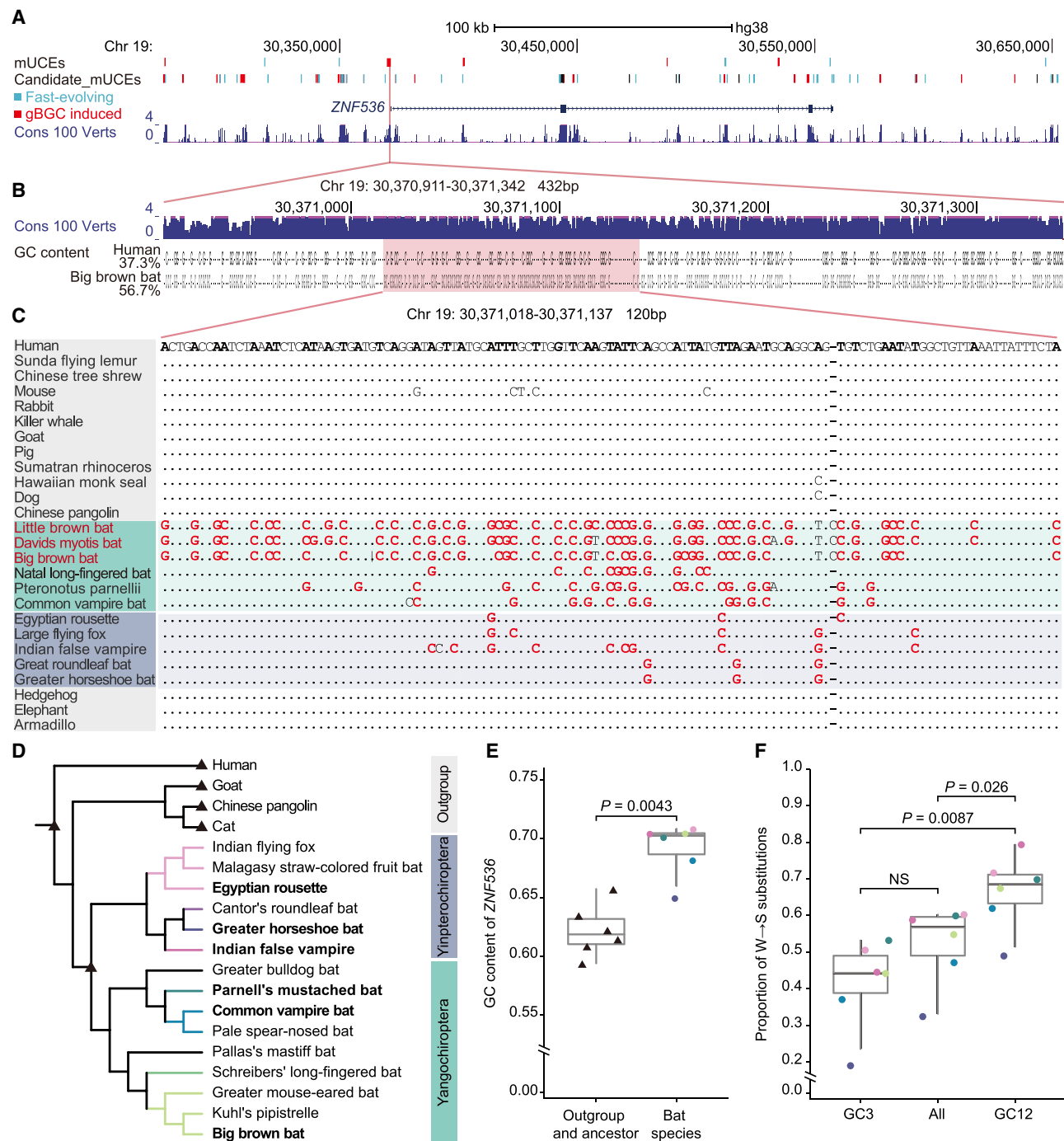


Figure 4. Example of gBGC-induced accelerated evolution of UCE in Chiroptera. (A) All mUCEs and candidate mUCEs in the region of *ZNF536* $\pm$ 100 kb. The elements in blue indicate fast-evolving elements, whereas gBGC-induced elements are shown in red. (B) The 432-bp mUCE.2036 is located in the *ZNF536* promoter region. Blue bars indicate conserved regions in 100 vertebrates. The GC content of mUCE.2036 was 37.3% and 56.7% in humans and big brown bats (Vespertilionidae), respectively. (C) Sequence alignment of a 120-bp subregion inside mUCE.2036. The shaded species names indicate species in Chiroptera, and the names in red indicate bats in Vespertilionidae. W $\rightarrow$ S substitutions are highlighted in red. Dots in the sequence alignment refer to bases that are identical to those in the human genome. (D) Phylogeny of 19 analyzed species. Four outgroup species and two ancestral nodes are marked with a black triangle, and selected bat species for subsequent analyses are marked in bold. The branches of different bat families are shown in different colors. (E) *ZNF536* GC content comparison between six bat species and outgroup and ancestor. The dot colors are the same as that of the corresponding bat family in D. (F) Comparison of proportion of W $\rightarrow$ S substitutions in different codon positions.

potentially change neurodevelopment, proliferation, and synaptic connectivity and thus potentially affect the neural network function that differentiated us from Neanderthals (Trujillo et al.

2021). To determine the in vivo effect of this cervid-specific GC-biased substitution pattern and further test our hypothesis, we used CRISPR-Cas9 gene editing to replace the 811-bp endogenous rat

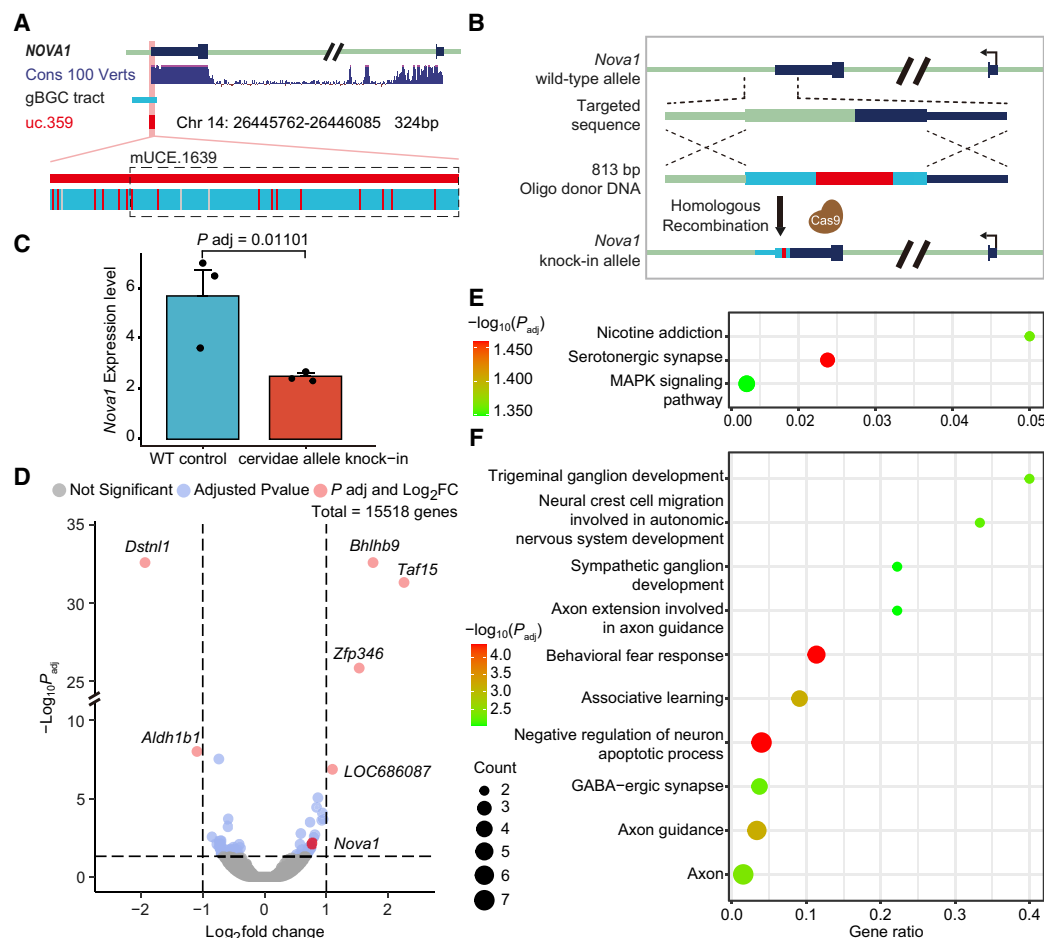


Figure 5. gBGC-induced functional alterations of UCEs in Cervidae. (A) uc.359 in Cervidae. Sixteen W → S cervid-specific substitutions are shown as red short lines. The black dotted rectangle is the 271-bp mUCE.1639 in our data set. (B) Gene editing abstract of the cervid-specific gBGC tract; the target region was replaced with an 811-bp cervid-specific gBGC tract. (C) Knock-in heterozygous rats showed a significant decrease of *Nova1* expression. (D) Volcano plot comparing wild-type and knock-in rats. Cutoff values of $|\text{fold change}| > 2$ and $\text{FDR} < 0.05$ were used to identify differentially expressed genes. (E) KEGG pathways of 72 significantly expressed genes in D. (F) Top 10 nervous system-related GO items of 72 significantly expressed genes in D.

sequence with a homologous 813-bp cervid sequence with the gBGC tract including uc.359 (Fig. 5B; Supplemental Table S15). Because no viable homozygous rats were obtained, we next examined whether *Nova1* expression in adult heterozygous rats was altered upon changes of the single allele's 3' UTR. RNA-seq on whole brain showed that a single cervid-specific allele resulted in significantly reduced expression of *Nova1* and other neural-related genes (adjusted P -value < 0.05) (Fig. 5C,D; Supplemental Table S16). KEGG and GO enrichment analyses showed that these genes were significantly enriched in neurodevelopment and other neuron-related processes (Fig. 5E,F; Supplemental Table S17); these findings were consistent with those of previous studies (Jensen et al. 2000; Ule et al. 2005; Trujillo et al. 2021).

Additionally, assigning RNA-seq reads of heterozygous rats to their genetic allele origin showed that cervid-specific alleles had lower expression than the rats' alleles (Supplemental Fig. S17). In total, 10/12 cervid-specific sites possessed significantly lower expression than the rats' sites (P -value < 0.05 ; Student's t -test), which further indicated potential down-regulation of cervid-specific mRNA expression. Together, these results indicated that the gBGC-induced alteration of UCEs, combined with selection, can moderately change phenotypes, at least at the expression level.

gBGC events may facilitate developmental regulatory divergence between mammals and birds

uc.359 provided an example of partial gBGC-induced acceleration conserved between lineages that could be an indicator of functional alterations. Next, we focused on more ancient gBGC events that may have directly affected eutherian evolution.

Using the same identification pipeline of gBGC, we compared mUCEs and aUCEs in reconstructed ancestral genomes of placental mammals and birds to infer ancient eutherian innovation of gBGC-induced UCEs. To make this inference more reliable, we used the platypus genome as an additional reference to conduct analysis. Using such an intermediate species was expected to reveal more specified functional elements during placental innovation. By intersecting platypus orthologs of mUCEs and aUCEs, we obtained 548 UCEs shared between mammals and birds that could be well recognized and may have been preserved during mammalian evolution (see Methods).

In total, seven of 548 UCEs were identified as ancient gBGC-induced UCEs. Notably, except one coding UCE, all ancient gBGC-induced UCEs showed clear enhancer-like signatures in at least two enhancer databases, which indicated their potential *cis*-regulatory

functions (Supplemental Table S18). Among them, mUCE.1304 showed perfect maintenance of an ancient gBGC event in eutherians. Notably, 22 W → S substitutions significantly (binomial test P -value = 2.91×10^{-3}) accumulated and were maintained in a 143-bp subregion (Fig. 6A–C). A motif scan identified potential alterations of predicted transcription factor binding sites induced by gBGC (Fig. 6D; Supplemental Table S19). To further investigate the potential functional impact of this gBGC-induced UCE, we used human and chicken orthologous sequences of mUCE.1304 (380 bp and 387 bp) (Supplemental Data S2) as the proxies of the corresponding mUCE and aUCE and performed dual-luciferase reporter gene experiments in both the human embryonic kidney 293T cell line and chicken embryonic fibroblast DF-1 cell line, respectively. The results showed that both sequences showed strong enhancer activity, with the aUCE showing significantly higher activity than the mUCE (P -value < 0.0001) (Fig. 6E). SEdb and SEA support that mUCE.1304 served as an enhancer involved in a superenhancer, one of which putative target genes encode an important transcription factor, paired box 2 (*PAX2*) (Chen et al. 2020; Wang et al. 2023), which has a critical role in the development of the central nervous system, kidney, urogenital tract, eyes, and inner ear (Torres et al. 1995, 1996; Favor et al. 1996; Urbánek et al. 1997; Bouchard et al. 2002; Stanke et al. 2010). These organs have modest but significant differences between mammals and birds. Therefore, it is highly possible that this ancient gBGC-induced UCE indicated developmental regulatory divergence and/or innovation emerged in eutherians. Briefly, our results indicated that these ancient gBGC-induced fast-evolving UCEs may have distinct roles in placental mammal development because they only underwent local gBGC events after differentiation from Monotremata and retained their ultraconservation under extreme selective pressure.

Discussion

In this study, we use nearly 200 genomes that covered all major mammalian and avian lineages to detect UCEs and explore the underlying roles of gBGC in shaping fast-evolving UCEs. Using our method, 2191 and 5938 credible UCEs were finally identified in mammals and birds, respectively. Further exploration of these UCEs showed their extreme conservation and potential functional importance. Consistent with previous studies, our study further provided multiple lines of evidence that the UCEs play a critical role in animal evolution: (1) More than half of UCEs were innovations produced in different evolutionary stages after divergence from the common ancestor with fish and were eventually “frozen” in birds and mammals (Bejerano et al. 2004); (2) compared with the RANDs, the strongly depleted divergent sites and a lower allelic frequency accounted for the majority of the lessened SNPs in UCEs indicated that UCEs were ultraselected regions rather than low-mutation regions during evolution (Drake et al. 2006; Katzman et al. 2007); (3) both the expression diversity index (CV) and tissue-specific index (τ) of UCE-adjacent genes were low, which indicated the genes were precisely regulated and tended to be widely expressed in the tissues; and (4) GO analysis of UCE-adjacent genes further supported the importance of UCEs by showing significant enrichment of basic biological processes and development, although there were different enrichment trends of neural-related processes and body development in mammals and birds, respectively (Bejerano et al. 2004; Seki et al. 2017). However, similar to the imperfect conservation of the 481 original UCEs that were identified by comparing 120 mammals (Hecker and Hiller 2020),

our results revealed that even the 48 most extreme mUCEs, which had absolutely identical sequences in all eight mammalian branches, also possessed mutations in specific species (Supplemental Fig. S4). In fact, as more mammalian species were included and used for comparison, more UCEs will not guarantee their ultraconservation. Thus, it is possible that nothing is unchangeable in genome evolution.

Although their extreme conservation and functional roles are not fully understood, these genome-wide distributed UCEs still represent a robust set of sequences that provide us with ideal materials for further exploring mutations and potential mechanisms that contribute to lineage-specific functional alteration. Focusing on fast-evolving UCEs, we found that gBGC played an important role in facilitating lineage-specific accelerated evolution of UCEs and shaped conspicuous local regions filled with exceeded GC substitutions. Further exploration revealed that groups and species with larger population sizes, smaller body sizes, and shorter generation times had relatively larger proportions of fast-evolving UCEs and gBGC-induced fast-evolving UCEs. Additionally, the distribution of UCEs affected by gBGC also varied among lineages and species. Such results appeared to reflect the general rules of gBGC (Romiguier et al. 2010; Weber et al. 2014).

Kostka et al. (2012) deduced that substitutions in many HARs cannot be explained by gBGC alone because the evidence of unusually high substitution rates exclusively emerged in humans, which indicated a possible selection role. In our study, gBGC detection extended even into those most extreme conserved regions in different lineages and provided several clues to further investigate the comprehensive impact of gBGC. Our *Nova1* gene-edited rats indeed provided some evidence regarding expression level because the cervid-specific allele expression was significantly lower than that of rats, which resulted in decreased *Nova1* expression. Although no homozygous rats were obtained, comparison between the wild and heterozygous rats still provided evidence that such cervid-specific GC-biased substitution patterns have a functional impact on expression pattern changes, which further supports the hypothesis that the fast-evolving UCEs induced by gBGC are highly possible and affect species or lineage adaptation, at least at the expression level (Necsulea and Kaessmann 2014). In mice, full knockout of *Nova1* leads to complete penetrance of postnatal lethal and severely impaired neural phenotypes (Jensen et al. 2000). Although no *Nova1* homozygous knockout rats were recorded, it is reasonable to speculate that similar or even more severe phenotypes can occur in rats considering their evolutionarily close relationship to mice.

Similarly, we also found that several ancient gBGC events occurred in mUCEs in earlier evolutionary periods of placental mammals. Potentially altered regulatory activities and the perfect maintenance of the conservation in mUCE.1304 indicate that gBGC events in UCEs may affect the adaptation and/or functional divergence of placental mammals and undergo natural selection. However, it is unclear whether the accumulated mutations were ultimately adaptive or a mixture of deleterious mutations driven by gBGC and compensatory mutations driven by positive selection that maintained their regulatory activities. Further investigations are needed to determine the evolutionary sources of mutations fixed in UCEs and the detailed functional impact of these altered elements. Nevertheless, both examples suggested that gBGC can contribute to the exploration of regions of the fitness landscape that would otherwise be unreachable owing to purifying selection, which prevents the fixation of deleterious intermediate genotypes between two fitness peaks.

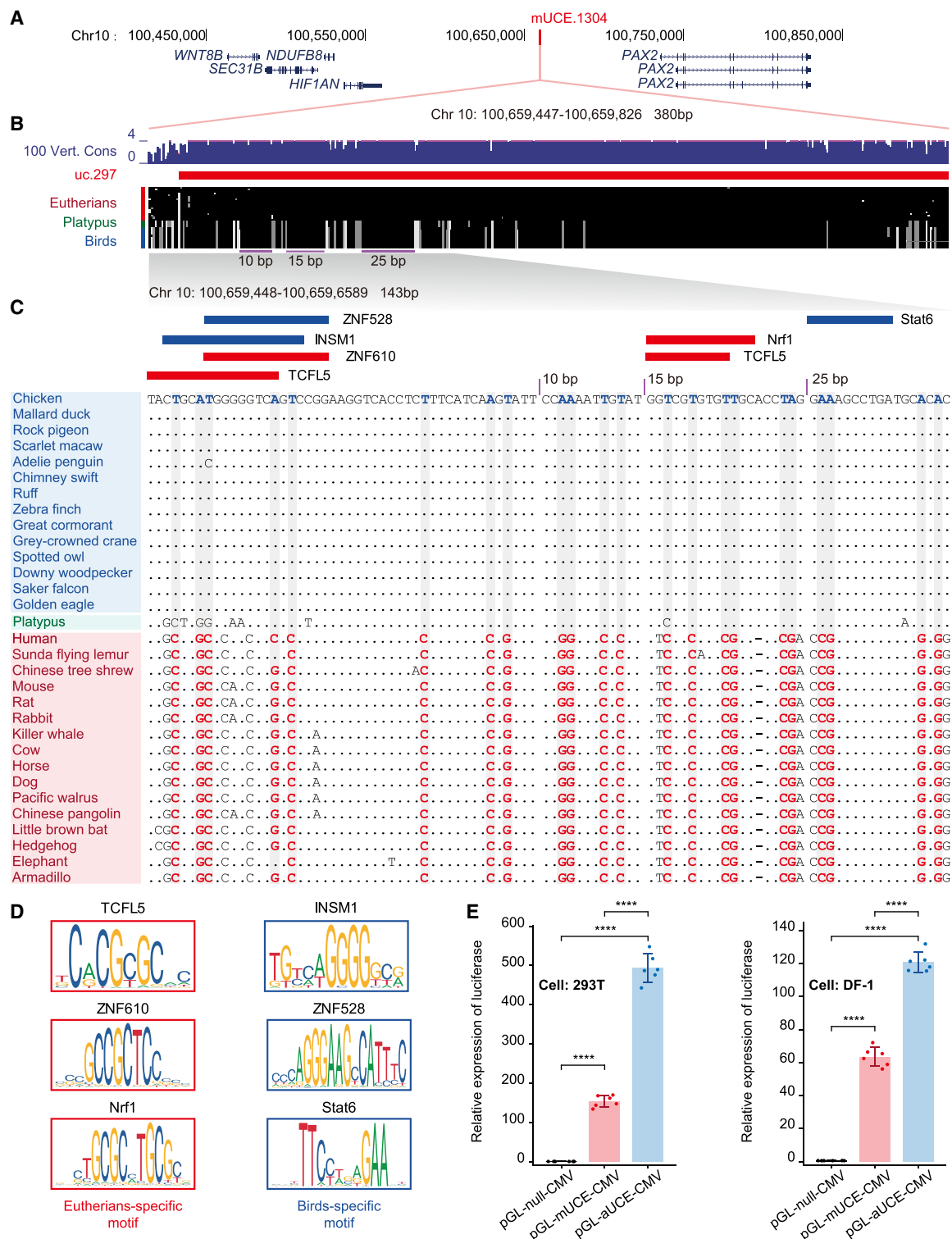


Figure 6. Example of an ancient gBGC-induced mUCE. (A) A 380-bp mUCE.1304 is located in the intergenic region close to *PAX2* and shows ultraconservation in eutherians and birds. (B) Alignment of mUCE.1304 shows that a gBGC event occurred in the eutherian ancestor and perfectly maintained ultraconservation during mammalian evolution. (C) Specific alignment of the 143-bp subregion. Dots in the sequence alignment refer to bases that are identical to those in the avian genome. (D) Specific changes in predicted transcription factor binding motifs owing to gBGC-induced alterations within the 143-bp subregion. The corresponding motif locations were shown in C. (E) Dual-luciferase reporter gene assay for the assessment of mUCE.1304 enhancer activity. Relative luciferase activity of reporter vectors containing mUCE (human ortholog of mUCE.1304) and aUCE (chicken ortholog of mUCE.1304) was measured in the human embryonic kidney 293T cell line (left) and chicken embryonic fibroblast DF-1 cell line (right). Six replications were performed for each experimental group. (****) P -value < 0.0001 .

Given the similar features of fixation along the genome, more detailed analyses should be performed for both genes and their proximal conserved noncoding regions before functional interpretation of positively selected genes or sites to exclude interference caused by gBGC, especially in cases such as chiropteran *ZNF536*. Our results combined with those of previous studies on HARs indicate that meticulous statistical and bioinformatic tests and functional studies are required for the appropriate interpretation of these adaptation-like accelerated elements (Galtier and Duret 2007; Kostka et al. 2012).

In conclusion, our comparative genomic approach identified an extended set of UCEs for both mammals and birds. These UCEs provide many new candidates for further exploring and decoding the key events in lineage evolution. Among these extreme conserved elements, the impact of gBGC was pervasive and seemingly strong enough to drive some modest alterations in specific lineages, especially in synergy with other evolutionary forces. Considering the different evolutionary background of the genomic regions where each element is located, it is necessary to take gBGC into account when testing for lineage-specific acceleration, especially when detecting positively selected elements.

Methods

Species genome information

The final collection of genomes used in this study contained 201 species, including 95 placental mammals, 94 birds, and 12 other vertebrate genomes (for the summarized detailed information, see Supplemental Tables S1, S2, S5). All of the genomes were downloaded from the National Center for Biotechnology Information (NCBI). The eutherian and avian genomes were used to identify the UCEs, whereas other vertebrate genomes were used to assess the conservation of UCEs.

Phylogeny tree construction

The fourfold degenerate sites were used to construct phylogenetic trees for both mammals and birds. First, we generated the pairwise sequence alignments across all selected mammalian and avian genomes by LAST (version last867; parameters “-m100 -E0.05”) (Kielbasa et al. 2011) using goat (ARS1) and chicken (GRCg6a) genomes as the reference, respectively. We then used MULTIZ (v11.2) (Blanchette et al. 2004) with the parameters “-O -all” to combine the pairwise alignments into a 95-way mammalian multiple sequence alignment and a 94-way avian multiple sequence alignment. Next, the fourfold degenerate sites of all the species were extracted from multiple sequence alignments according to the gene annotations of the reference genome. Fourfold degenerate site sequences were trimmed by discarding the poorly aligned positions and divergent regions. Finally, the phylogeny tree was constructed by IQ-TREE (v1.6.7) with the “-bb 1000, -m TEST” parameter, which means 1000-bootstrap and standard-model selection followed by tree inference (Nguyen et al. 2015).

An approach to detect mUCEs and aUCEs

To detect mUCEs and aUCEs, we used 95 eutherian and 94 avian genomes (for list, see Supplemental Tables S1, S2) encompassing all orders of placental mammals and birds, respectively. Here, taking the detection of mUCEs as an example, the detailed method is shown as follows. Scripts used in UCE identification are publicly available (see Data access).

First, we constructed the mammalian ancestral genome using 10 high-quality genomes representing 10 orders of mammals (Pholidota: Malayan-pangolin; Carnivora: cat; Perissodactyla: horse; Chiroptera: Egyptian rousette; Cetartiodactyla: goat; Eulipotyphla: star-nosed mole; Rodentia: mouse; Primates: human; Proboscidea [representation of Afrotheria]: elephant; Cingulata [representation of Xenarthra]: armadillo) (Fig. 1A). In brief, to reduce the impact of reference bias, these genomes were used to implement a whole-genome alignment using the armadillo (split from the rest of the species ~100 million years ago) as the reference genome. Pairwise alignments were generated using LAST (version last867; parameters “-m100 -E0.05”) (Kielbasa et al. 2011), and multiple alignments were generated using MULTIZ (v11.2) (Blanchette et al. 2004) with the parameters “-O -all”. Then the neutral phylogenetic model used for constructing the ancestor genome was estimated using fourfold degenerate sites, which were extracted from above multiple alignment by phyloFit (Hubisz et al. 2011). Then, the ancestor genome was constructed by prequel (Hubisz et al. 2011) with the neutral phylogenetic model. Next, all 95 mammalian genomes were simulated to 50-mers sequences (10-bp bootstrap) and aligned with the mammalian ancestor genome to obtain the unchanged sequences using Bowtie 2 (v2.2.8) (Langmead and Salzberg 2012) with default parameters.

Then, sequences that could be absolutely aligned to the ancestral genome without any mismatch were extracted from the ancestral genome as the mammalian candidate conserved regions. Next, the candidate conserved regions were merged to obtain as long as possible candidate UCEs (≥ 100 bp at least) (Fig. 1B). The length threshold of human-rodents UCEs was 200 bp and required absolute conservation across the mouse, rat, and human genomes (Bejerano et al. 2004), which were within ~75 million years of evolution (Kumar et al. 2017). To obtain more accurate UCEs, we aligned the candidate UCEs again to all 95 mammals using BLAST v2.11.0 (Camacho et al. 2009). Considering many more species were included in our study, the candidate UCEs, which were 100% identical within at least one branch (thus, three or more species) and with at least 80% identity across all branches (~100 million years of evolution distance) (Kumar et al. 2017), were defined as final UCEs. To again reduce the impact of reference bias, we did not include the armadillo genome in any of the branches. Finally, we identified 20,319 mUCEs >100 bp and 2191 mUCEs >200 bp.

The same pipeline was used to identify the aUCEs (Fig. 1). Ten avian high-quality genomes were selected to construct avian ancestral genome: Gruiformes, gray crowned crane; Falconiformes, golden eagle; Pelecaniformes, Galapagos flightless cormorant; Columbiformes, rock pigeon; Apodiformes, chimney swift; Passeriformes, tawny-bellied seedeater; Charadriiformes, ruff; Psittaciformes, budgerigar; Galliformes, chicken; and Struthioniformes, ostrich. Here, we used the ostrich (split from the rest of the species ~100 million years ago) genome as a reference to obtain the whole-genome alignment. Finally, 30,154 aUCEs with a length >100 bp and 5938 aUCEs >200 bp were identified. Please note that the avian phylogeny generated by trimmed fourfold degenerate site sequences was not fully consistent with previous findings (Jarvis et al. 2014; Prum et al. 2015) owing to pervasive incomplete lineage sorting in the deep branches of the species tree. However, this incongruence should not have a negative impact upon the accuracy of aUCE identification, especially in situations in which closely related species were assigned to different branches. Likewise, it should not affect the phylogenetic relationships and subsequent analyses of the 10 test avian lineages.

Genomic distribution annotation of UCEs

We used ANNOVAR (Wang et al. 2010) and human and chicken genome annotation (GRCh38.p12 and GRCg6a) to classify the UCEs into six nonoverlapping groups: exonic, intronic, UTR, intergenic, ncRNA exonic, and ncRNA intronic.

Conservation analysis of UCEs

To detect the conservation of UCEs along vertebrate evolution, we aligned mUCEs and aUCEs to 12 evolutionarily distant vertebrate genomes (for list, see [Supplemental Table S5](#)) using BLAST v2.11.0 (Camacho et al. 2009) with the parameter settings “-task blastn, -outfmt 6, -max_target_seqs 1, -max_hsps 1.”

Identification of fast-evolving UCEs

To detect the fast-evolving UCEs in selected mammalian and avian lineages, we first estimated the neutral model of evolution using the fourfold degenerate sites by phyloFit in PHAST package v1.4 (Hubisz et al. 2011). Then, the fast-evolving UCEs of each test lineage were detected by phyloP in PHAST package v1.4 (Hubisz et al. 2011) using CONACC mode and the LRT method. The final fast-evolving UCEs are a statistically significant departure from the neutral model with $\text{altscale} > 1$ and adjusted P -value < 0.05 (Benjamini–Hochberg algorithm).

gBGC detection

To evaluate the gBGC among fast-evolving UCEs, we first used phastBias program (Capra et al. 2013) using default parameters to output informative gBGC regions and gBGC tracts for all test lineages. Here, informative gBGC regions were defined as regions where $>50\%$ of the sites were informative for gBGC (computed by phastBias), whereas gBGC tracts were defined as regions with a posterior probability of being in a gBGC state > 0.5 . We intersected informative gBGC regions with fast-evolving UCEs, and if the overlapping intervals are longer than half of the original length of UCEs, such UCEs would be identified as the informative gBGC regions. Similarly, if fast-evolving UCEs overlapped with gBGC tracts, they would be directly identified as informative gBGC UCEs because of their high confidence for gBGC. This prior detection would give a rough estimation for both the upper and lower limits of gBGC-induced fast-evolving UCEs.

Then, we used traditional summary statistics to infer the local influence of gBGC. The substitution numbers of each UCE were counted according to the mutation direction including (1) $W \rightarrow S$ and (2) $S \rightarrow W$. We define $P1$ as the percentage of $S \rightarrow W$ substitution type and $P2$ as the percentage of $W \rightarrow S$ substitution type. As the mutations are expected to occur randomly, the null hypothesis is the $P1 = P2 = 0.5$. For the UCEs with gBGC-induced acceleration, $P2 - P1$ is close to one. Here, based on the intrinsic extreme conservation of UCEs, we used an arbitrary but strict cutoff that $(P2 - P1) > 0.9$ to infer the true gBGC-induced acceleration. The Storer–Kim test (Storer and Kim 1990) was used to test the null hypothesis that $(P2 - P1) = 0$. The P -value was corrected using the Benjamini–Hochberg algorithm to identify gBGC-induced fast-evolving UCEs (adjusted P -value < 0.05).

Comparing SNP frequency and the counts of divergent sites in UCEs and RANDs

For mUCEs, we accessed the 1000 Genomes Project (Byrsk-Bishop et al. 2022) high-coverage genotype data (3202 human samples) from https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20220422_3202_phased_SNV_INDEL_SV. We used human SNP data and removed the sites with

more than three alleles or singletons for the following analyses. Finally, 61,477,086 human SNPs were used to perform subsequent analyses. We first used all SNPs to calculate the SNP enrichment in UCEs and RANDs. We then partitioned SNPs into three groups—(1) derived allele frequency (DAF) ≤ 0.01 , (2) $0.01 < \text{DAF} \leq 0.05$, and (3) $\text{DAF} > 0.05$ —and performed the same analyses. The divergent sites were generated from the pairwise alignment between the chimpanzee and human genomes by LAST (version last867; parameters -m100 -E0.05) (Kielbasa et al. 2011), and we extracted the location of the divergent sites from the BAM file. For aUCEs, we performed same analyses using chicken SNPs generated from 928 chicken genomes (Fu et al. 2022) and divergent sites between chicken and turkey. Finally, 21,672,488 chicken SNPs were used to perform subsequent analyses.

We obtained RANDs using BEDTools shuffle (Quinlan and Hall 2010). The coverage of the SNPs and divergent sites between UCEs and RANDs were counted by BEDTools coverage (Quinlan and Hall 2010). The P -value was calculated by the Wilcoxon rank-sum test.

Transcriptomic analysis

We downloaded 55 human transcriptomes of 11 tissues and 62 chicken transcriptomes of 17 tissues (for list, see [Supplemental Table S6](#)) to perform the following analysis. Using human (GRCh38.p12) and chicken (GRCg6a) genomes as references, we filtered low-quality reads using Trimmomatic v0.36 (Bolger et al. 2014) with the following parameters: LEADING:3, TRAILING:3, SLIDINGWINDOW:4:15, and MINLEN:40. The remaining high-quality RNA-seq clean reads were aligned to the corresponding reference genomes by the HISAT2 v2.0.3 (Kim et al. 2015) program with default parameters. The gene expression levels of the detected tissues were quantified (in fragments per kilobase of transcript per million mapped reads [FPKM]) by StringTie (Pertea et al. 2015) based on the corresponding transcript annotation.

Calculation of τ index

To calculate tissue-specific expression index, τ , which is the best indicator of tissue-specific expression according to a comparative review (Kryuchkova-Mostacci and Robinson-Rechavi 2017), τ is defined as follows:

$$\tau = \frac{\sum_{i=1}^n (1 - \hat{x}_i)}{n - 1}; \hat{x}_i = \frac{x_i}{\max_{1 \leq i \leq n} (x_i)}.$$

The value of τ is between zero and one. Values close to one indicate tissue-specifically expressed genes, and values close to zero indicate ubiquitously expressed genes.

Calculation of CV value

Gene expression diversity refers to the degree of variation in the expression levels of a specific gene among different samples or tissues and can be reflected by the coefficient of variation (CV) of gene expression levels (Bellucci et al. 2014). We first calculated the CV value of each gene for each tissue using the corresponding multiple transcriptome data. The CV value was calculated as the ratio between the standard deviation (SD) and mean of gene expression levels (in FPKM) of each selected tissue obtained from the transcriptomic analysis. The gene expression diversity of each tissue in humans and chickens was calculated separately. Last, the mean CV of all selected tissues was used to represent the final CV of each gene.

GO and KEGG enrichment analysis of adjacent genes of UCEs

We annotated adjacent genes of mUCEs and aUCEs using ANNOVAR (Wang et al. 2010), respectively. Then, GO enrichment analyses of UCE-adjacent genes were performed using clusterProfiler (Wu et al. 2021). We used the simplify function in the clusterProfiler package with default parameters (cutoff = 0.7, by = “p.adjust”, select_fun = min, measure = “Wang”) to generate nonredundant GO terms as the input files for calculating the final similarity index in GOSemSim (Yu et al. 2010).

Life-history trait analysis

We performed nonparametric Spearman’s correlation tests between life-history traits (body mass and generation time) and the number of (1) fast-evolving UCEs and (2) gBGC-induced fast-evolving UCEs for each species (excluding armadillo and ostrich). Data on mammalian life-history traits were collected from the AnAge database (de Magalhães and Costa 2009), from the PanTHERIA database (Jones et al. 2009), and from Pacifici et al. (2013) and Ernest (2003). We collected avian body mass data from the AVONET database (Tobias et al. 2022) and generation time from Bird et al. (2020).

Additional tests were conducted to examine the relationships between the number of fast-evolving UCEs and three other confounders: (1) genome-wide neutral substitution rate, (2) assembly size, and (3) the number of species included in each test lineage. The genome-wide neutral substitution rate for each analyzed species was calculated based on the fourfold degenerate sites that were used to build the phylogenetic tree. All the data used for these analyses can be found in Supplemental Tables S1 and S2.

Interspecific synteny filtering

To accurately find those fast-evolving UCEs that kept good synteny with their target genes during chiropteran evolution, we filtered UCEs by comparing their corresponding genomic locations in three genomes, including the reconstructed mammalian ancestor, human, and goat. Only those fast-evolving UCEs that were found in same chromosomes in all three genomes were finally defined as potential regulatory elements to target genes. Then, we used BEDTools (Quinlan and Hall 2010) to intersect each UCE and gene ± 1 -Mb region of reference genome (ARS1) to obtain the genes harboring the highest amount of fast-evolving UCEs.

The d_N/d_S analysis of *ZNF536*

We selected 69 mammalian species used in mUCE identification to perform d_N/d_S analysis of *ZNF536*. We extracted *ZNF536* alignment (longest transcript of *ZNF536* in reference assembly) from our 95-way mammalian multiple sequence alignments. We estimated branch-specific d_N/d_S , or omega (ω), across the mammalian phylogeny, using the CODEML in the PAML v4.9e (Yang 2007). We used the branch model with freely varying omega (model = 1, NsSites = 0) to infer d_N/d_S across all branches separately (Supplemental Data S3).

Nineteen-way *ZNF536* coding sequence alignment

To explore the substitution pattern of *ZNF536* in bat evolution, we selected 19 high-quality genomes (for list, see Supplemental Table S12) that contain complete *ZNF536* coding sequence. This set includes 15 bat species, and six of them were used in UCE identification. We first used the canonical conserved transcript of human *ZNF536* (ENST00000355537.4) as the reference to obtain pairwise alignments by TOGA (Kirilenko et al. 2023). Then, we generated a 19-way *ZNF536* coding sequence alignment using MACSE V2

(Ranwez et al. 2018) with default parameter settings and removed gap sites for subsequent analyses (Supplemental Data S4).

Further investigations on branch-specific d_N/d_S were performed using the same method as mentioned above. Positively selected sites in each selected bat lineage were inferred by branch-site model implemented in CODEML in the PAML v4.9e (Yang 2007).

Generation of Cervidae-specific *Noval* rats

To determine the in vivo effects of Cervidae-specific GC-biased substitutions in the 3' UTR of *Noval*, an 813-bp homologous cervid DNA segment harboring predicted gBGC tract (embracing uc.359, total identities was 90.6%) was introduced into rats to replace the original allele segment by the CRISPR–Cas9 system as follows. First, single guide RNA (sgRNA) was constructed using a pUC57-sgRNA expression vector (Addgene 51132) and oligonucleotide sequences (listed in Supplemental Table S20). Next, the constructed sgRNA vector was in vitro transcribed into injectable sgRNA using a MEGAShortscript T7 kit (Ambion AM1354). Cas9 was amplified from plasmid pST1374-NLS-flag-linker-Cas9 (Addgene 44758), and the resulting Cas9-mRNA was in vitro transcribed using the mMESSAGE mMACHINE T7 kit (Ambion AM1345). The sgRNA and Cas9-mRNA were purified using a MEGAclear kit (Ambion AM1908). Fertilized eggs of Sprague Dawley (SD) rats were injected with a mixture of Cas9-mRNA, sgRNA, and homologous DNA segment to obtain F₀ knock-in rats. Genomic DNA was then extracted from the tails of 7-d-old rats for further polymerase chain reaction (PCR) genotyping (for list of primers, see Supplemental Table S21). Last, F₀ knock-in rats were mated with wild-type rats to get enough heterozygotes for consequence experiments.

We confirmed the correct gene-editing heterozygotes using Sanger sequencing (Supplemental Fig. S15B). DNA extraction from rat liver tissues was performed using the TaKaRa MiniBEST universal genomic DNA extraction kit ver.5.0 (TaKaRa 9765) according to the manufacturer’s instructions. Two pairs of identification primers were designed (F1: TGCAACCAATTAAAGAAC; R1: CTGAACAGCCATCGTCAC; F2: TGTCTCAGAGTCAGCACCGC; R2: GTTACAGATGTGATGGGAAGCTG), and PCR was performed with TransTaq DNA polymerase high fidelity (Transgen AP131). PCR products were cloned into T-vector pMD19 (Takara 3271) via DNA ligation kit (Takara 6022) to construct the final vector. All constructs were verified by Sanger sequencing (Tsingke Biotechnology).

RNA-seq of rat whole brain

Six whole-brain samples were obtained from three 2-mo-old wild-type male rats and three 2-mo-old gene-editing male heterozygotes. RNA was isolated according to the TRIzol (Invitrogen) protocol, and 1.5 μ g per sample was used as the input material for RNA sample preparations. Sequencing libraries were generated and sequenced using Illumina HiSeq X Ten, and 150-bp PE reads were generated.

The same transcriptomic analysis methods were performed as described in the section Transcriptomic Analysis. Differentially expressed genes (DEGs) between two sets of samples were identified as follows. Nonnormalized read counts for all detected genes were acquired by StringTie, and a read count table was generated by the Python script “prepDE.py” in the StringTie package. Then, the DEGs were identified by negative binomial generalized linear models implemented in DESeq2 v1.20.0 (Love et al. 2014). We identified 72 DEGs with an adjusted *P*-value < 0.05 (Benjamini–Hochberg algorithm) (Fig. 5D; Supplemental Table S16). Subsequent GO and KEGG enrichment analyses were

performed using online KOBAS-i (Bu et al. 2021; <http://kobas.cbi.pku.edu.cn/>).

Comparison of allele-specific expression

To compare the expression level of rat-specific and cervid-specific allele, we used RNA-seq data generated from heterozygous rats and assigned RNA-seq reads to their genetic allele origin following a previous method (Wang et al. 2019a). We calculated the rat and cervid allelic sites in all mapped paired-end reads in the target region. We then compared the counts of the 12 fixed divergent sites to reveal the differential expression level of *Nova1* between two alleles.

Identification of gBGC-induced UCEs in eutherian evolution

To identify ancient gBGC events that may affect evolution of placental mammals, we first used BLAST v2.11.0 (Camacho et al. 2009) to obtain corresponding orthologs of mUCEs and aUCEs in the platypus genome, respectively. The parameters were set as “-task blastn, -outfmt 6, -max_target_seqs 1, -max_hsps 1.” Then, we intersected platypus orthologs of mUCEs and aUCEs to obtain the homologous sequences for which overlapping intervals are >100 bp. The resulting homologous sequences and their corresponding mUCEs and aUCEs were again compared using same identification pipeline of gBGC. The sequences, which showed significant W → S substitutions in eutherian–platypus comparison but no evident W → S substitutions in bird–platypus comparison, were finally identified as gBGC-induced UCEs in eutherian evolution.

Regulatory activity analysis of eutherian gBGC-induced UCEs

To discover potential regulatory activities of eutherian gBGC-induced UCEs, we queried the genome location of these seven UCEs in ENCODE, scEnhancer, and SEA 3.0 to find possible records. For different input requirements (which mainly depend on the genome version), we used UCSC liftOver (Gonzalez et al. 2021) to convert the genome locations.

Motif analysis

Orthologous aligned mUCE.1304 sequences from eutherians (380 bp) and birds (387 bp) were scanned for all putative transcription factor binding sites using FIMO (Grant et al. 2011) and available nonredundant position frequency matrices from the JASPAR vertebrate CORE collection (Castro-Mondragon et al. 2022), respectively. Gaps were removed from the multispecies alignment.

Dual-luciferase reporter experiment

HEK293T cell lines were cultured in Dulbecco's Modified Eagle Medium, and all media were supplemented with 10% heat-inactivated fetal bovine serum (Gibco). DF-1 chicken embryo fibroblasts were cultured in Dulbecco's Modified Eagle Medium, and media were supplemented with 10% heat-inactivated fetal bovine serum (Gibco) and chicken serum (Solarbio). Both cell lines were maintained in a 5% CO₂-humidified atmosphere at 37°C.

The CMV promoter was amplified via PCR from the p3 × FLAG-CMV-9 expression vector (Sigma-Aldrich) using primers that contained the requisite recombination sites. The resulting PCR product was purified and subsequently inserted into the HindIII and BglII sites of pGL4.10. The two UCE sequences (380-bp mUCE and 387-bp aUCE) were synthesized (Beijing Tsingke Biotech) and subsequently subcloned into the CMV-Luc vector.

HEK293T cells were seeded in the wells of a 24-well plate at a density of 1×10^5 cells per well 1 d before transfection; 1.0 µg of

plasmids (0.9 µg for luciferase reporter plasmid and 0.1 µg for *Renilla*-luciferase plasmid [pRL-TK]) was cotransfected according to the protocol of the FuGENE HD transfection reagent (Promega). DF-1 cells were seeded in the wells of a 12-well plate at a density of 3×10^5 cells per well 1 d before transfection; 2.0 µg of plasmids (1.8 µg for luciferase reporter plasmid and 0.2 µg for pRL-TK) was cotransfected according to the protocol of the FuGENE HD transfection reagent (Promega). A PRL-TK was cotransfected to control for transfection efficiency. Cell lysates were collected 48 h post-transfection and prepared for luciferase activity analysis using the double-luciferase reporter assay kit (TransGen) following the manufacturer's instructions. Relative luciferase activities were expressed as the ratio of the luciferase value to the *Renilla* value. Six replications were performed for each experimental group.

Data access

The raw RNA-seq data generated in this study have been submitted to the NCBI BioProject database (<https://www.ncbi.nlm.nih.gov/bioproject/>) under accession number PRJNA909437 and to the CNGB Sequence Archive (CNSA) of the China National GeneBank DataBase (CNGBdb; <https://db.cngb.org>) under accession number CNP0003788. The workflows and associated codes and scripts for analyses described in this study are available at GitHub (<https://github.com/Anguo-Liu/gBGC-drives-accelerated-evolution-of-UCes>) and as Supplemental Code.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

This project was supported by the National Key R&D Program of China (2021YFF1001000) and the Postdoctoral Innovative Talents Support Program of China (BX20200282) to Y.W. We thank the high-performance computing platform of Northwest A&F University (NWAUFU) and the Hefei Advanced Computing Center for providing computing resources. We thank Mallory Eckstut, PhD, for editing the English text of a draft of this manuscript.

Author contributions: Y.W. conceived the project and designed research. A.G.L., N.N.W., and G.X.X. performed the majority of analysis with contributions from Y.L., X.X.Y., X.M.L., Z.H.L., F.X.M., M.L.D., W.H.C., and N.G.M. Z.L.Z., J.Y., and Y.P.G. contributed to the experiments. A.G.L. and N.N.W. drafted the manuscript with input from all authors, whereas Y.W., Y.P.G., and Y.J. revised the manuscript.

References

- Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. 2004. Ultraconserved elements in the human genome. *Science* **304**: 1321–1325. doi:10.1126/science.1098119
- Bellucci E, Bitocchi E, Ferrarini A, Benazzo A, Biagetti E, Klie S, Minio A, Rau D, Rodriguez M, Panziera A, et al. 2014. Decreased nucleotide and expression diversity and modified coexpression patterns characterize domestication in the common bean. *Plant Cell* **26**: 1901–1912. doi:10.1105/tpc.114.124040
- Bird JP, Martin R, Akçakaya HR, Gilroy J, Burfield IJ, Garnett ST, Symes A, Taylor J, Şekerioğlu CH, Butchart SHM. 2020. Generation lengths of the world's birds and their implications for extinction risk. *Conserv Biol* **34**: 1252–1261. doi:10.1111/cobi.13486
- Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* **14**: 708–715. doi:10.1101/gr.1933104

- Boffelli D, Nobrega MA, Rubin EM. 2004. Comparative genomics at the vertebrate extremes. *Nat Rev Genet* **5**: 456–465. doi:10.1038/nrg1350
- Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120. doi:10.1093/bioinformatics/btu170
- Booker BM, Friedrich T, Mason MK, VanderMeer JE, Zhao J, Eckalbar WL, Logan M, Illing N, Pollard KS, Ahituv N. 2016. Bat accelerated regions identify a bat forelimb specific enhancer in the HoxD locus. *PLoS Genet* **12**: e1005738. doi:10.1371/journal.pgen.1005738
- Borges R, Szöllösi GJ, Kosiol C. 2019. Quantifying GC-biased gene conversion in great ape genomes using polymorphism-aware models. *Genetics* **212**: 1321–1336. doi:10.1534/genetics.119.302074
- Bouchard M, Souabni A, Mandler M, Neubüser A, Busslinger M. 2002. Nephric lineage specification by Pax2 and Pax8. *Genes Dev* **16**: 2958–2970. doi:10.1101/gad.240102
- Braconi C, Valeri N, Kogure T, Gasparini P, Huang N, Nuovo GJ, Terracciano L, Croce CM, Patel T. 2011. Expression and functional role of a transcribed noncoding RNA with an ultraconserved element in hepatocellular carcinoma. *Proc Natl Acad Sci* **108**: 786–791. doi:10.1073/pnas.1011098108
- Bu D, Luo H, Huo P, Wang Z, Zhang S, He Z, Wu Y, Zhao L, Liu J, Guo J, et al. 2021. KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis. *Nucleic Acids Res* **49**: W317–W325. doi:10.1093/nar/gkab447
- Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, Corvelo A, Clarke WE, Musunuri R, Nagulapalli K, et al. 2022. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**: 3426–3440.e19. doi:10.1016/j.cell.2022.08.004
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**: 421. doi:10.1186/1471-2105-10-421
- Capra JA, Hubisz MJ, Kostka D, Pollard KS, Siepel A. 2013. A model-based analysis of GC-biased gene conversion in the human and chimpanzee genomes. *PLoS Genet* **9**: e1003684. doi:10.1371/journal.pgen.1003684
- Castro-Mondragon JA, Riudavets-Puig R, Rauluseviciute I, Lemma RB, Turchi L, Blanc-Mathieu R, Lucas J, Boddie P, Khan A, Perez NM, et al. 2022. JASPAR 2022: the ninth release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **50**: D165–D173. doi:10.1093/nar/gkab1113
- Chen CT, Wang JC, Cohen BA. 2007. The strength of selection on ultraconserved elements in the human genome. *Am J Hum Genet* **80**: 692–704. doi:10.1086/513149
- Chen C, Zhou D, Gu Y, Wang C, Zhang M, Lin X, Xing J, Wang H, Zhang Y. 2020. SEA version 3.0: a comprehensive extension and update of the super-enhancer archive. *Nucleic Acids Res* **48**: D198–D203.
- Christmas MJ, Kaplow IM, Genereux DP, Dong MX, Hughes GM, Li X, Sullivan PF, Hindle AG, Andrews G, Armstrong JC, et al. 2023. Evolutionary constraint and innovation across hundreds of placental mammals. *Science* **380**: eabn3943. doi:10.1126/science.abn3943
- de la Calle-Mustienes E, Feijóo CG, Manzanares M, Tena JJ, Rodríguez-Seguel E, Letizia A, Allende ML, Gómez-Skarmeta JL. 2005. A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate *Iroquois* cluster gene deserts. *Genome Res* **15**: 1061–1072. doi:10.1101/gr.4004805
- de Magalhães JP, Costa J. 2009. A database of vertebrate longevity records and their relation to other life-history traits. *J Evol Biol* **22**: 1770–1774. doi:10.1111/j.1420-9101.2009.01783.x
- Dickel DE, Ypsilanti AR, Pla R, Zhu Y, Barozzi I, Mannion BJ, Khin YS, Fukuda-Yuzawa Y, Plajzer-Frick I, Pickle CS, et al. 2018. Ultraconserved enhancers are required for normal development. *Cell* **172**: 491–499.e15. doi:10.1016/j.cell.2017.12.017
- Drake JA, Bird C, Nemesh J, Thomas DJ, Newton-Cheh C, Reymond A, Excoffier L, Attar H, Antonarakis SE, Dermitzakis ET, et al. 2006. Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nat Genet* **38**: 223–227. doi:10.1038/ng1710
- Duret L, Galtier N. 2009a. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet* **10**: 285–311. doi:10.1146/annurev-genom-082908-150001
- Duret L, Galtier N. 2009b. Comment on 'Human-specific gain of function in a developmental enhancer.' *Science* **323**: 714. doi:10.1126/science.1165848
- Dutrow EV, Emera D, Yim K, Uebbing S, Kocher AA, Krenzer M, Nottoli T, Burkhardt DB, Krishnaswamy S, Louvi A, et al. 2022. Modeling uniquely human gene regulatory function via targeted humanization of the mouse genome. *Nat Commun* **13**: 304. doi:10.1038/s41467-021-27899-w
- Ernest SKM. 2003. Life history characteristics of placental nonvolant mammals. *Ecology* **84**: 3402. doi:10.1890/02-9002
- Favor J, Sandulache R, Neuhäuser-Klaus A, Pretsch W, Chatterjee B, Senft E, Wurst W, Blanquet V, Grimes P, Spörle R, et al. 1996. The mouse *Pax2*^{1Neu} mutation is identical to a human *PAX2* mutation in a family with renal-coloboma syndrome and results in developmental defects of the brain, ear, eye, and kidney. *Proc Natl Acad Sci* **93**: 13870–13875. doi:10.1073/pnas.93.24.13870
- Feigin CY, Newton AH, Pask AJ. 2019. Widespread cis-regulatory convergence between the extinct Tasmanian tiger and gray wolf. *Genome Res* **29**: 1648–1658. doi:10.1101/gr.244251.118
- Ferris E, Abegglen LM, Schiffman JD, Gregg C. 2018. Accelerated evolution in distinctive species reveals candidate elements for clinically relevant traits, including mutation and cancer resistance. *Cell Rep* **22**: 2742–2755. doi:10.1016/j.celrep.2018.02.008
- Figuet E, Ballenghien M, Romiguier J, Galtier N. 2014. Biased gene conversion and GC-content evolution in the coding sequences of reptiles and vertebrates. *Genome Biol Evol* **7**: 240–250. doi:10.1093/gbe/evu277
- Fu W, Wang R, Xu N, Wang J, Li R, Asadollahpour Nanaei H, Nie Q, Zhao X, Han J, Yang N, et al. 2022. Galbase: a comprehensive repository for integrating chicken multi-omics data. *BMC Genomics* **23**: 364. doi:10.1186/s12864-022-08598-2
- Galtier N, Duret L. 2007. Adaptation or biased gene conversion? extending the null hypothesis of molecular evolution. *Trends Genet* **23**: 273–277. doi:10.1016/j.tig.2007.03.011
- Galtier N, Duret L, Glémin S, Ranwez V. 2009. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Trends Genet* **25**: 1–5. doi:10.1016/j.tig.2008.10.011
- Glémin S. 2010. Surprising fitness consequences of GC-biased gene conversion: I. mutation load and inbreeding depression. *Genetics* **185**: 939–959. doi:10.1534/genetics.110.116368
- Glémin S, Arndt PF, Messer PW, Petrov D, Galtier N, Duret L. 2015. Quantification of GC-biased gene conversion in the human genome. *Genome Res* **25**: 1215–1228. doi:10.1101/gr.185488.114
- 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 Res* **49**: D1046–D1057. doi:10.1093/nar/gkaa1070
- Gotea V, Elnitski L. 2014. Ascertain regions affected by GC-biased gene conversion through weak-to-strong mutational hotspots. *Genomics* **103**: 349–356. doi:10.1016/j.ygeno.2014.04.001
- Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**: 1017–1018. doi:10.1093/bioinformatics/btr064
- Hecker N, Hiller M. 2020. A genome alignment of 120 mammals highlights ultraconserved element variability and placenta-associated enhancers. *Gigascience* **9**: giz159. doi:10.1093/gigascience/giz159
- Holloway AK, Bruneau BG, Sukonnik T, Rubenstein JL, Pollard KS. 2016. Accelerated evolution of enhancer hotspots in the mammal ancestor. *Mol Biol Evol* **33**: 1008–1018. doi:10.1093/molbev/msv344
- Hubisz MJ, Pollard KS, Siepel A. 2011. PHAST and RPHAST: phylogenetic analysis with space/time models. *Brief Bioinformatics* **12**: 41–51. doi:10.1093/bib/bbq072
- Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SY, Faircloth BC, Nabholz B, Howard JT, et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* **346**: 1320–1331. doi:10.1126/science.1253451
- Jebb D, Huang Z, Pippel M, Hughes GM, Lavrichenko K, Devanna P, Winkler S, Jermin LS, Skirmuntt EC, Katourakis A, et al. 2020. Six reference-quality genomes reveal evolution of bat adaptations. *Nature* **583**: 578–584. doi:10.1038/s41586-020-2486-3
- Jensen KB, Dredge BK, Stefani G, Zhong R, Buckanovich RJ, Okano HJ, Yang YYL, Darnell RB. 2000. Nova-1 regulates neuron-specific alternative splicing and is essential for neuronal viability. *Neuron* **25**: 359–371. doi:10.1016/S0896-6273(00)80900-9
- Jones KE, Bielby J, Cardillo M, Fritz SA, O'Dell J, Orme CDL, Safi K, Sechrest W, Boakes EH, Carbone C, et al. 2009. PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology* **90**: 2648. doi:10.1890/08-1494.1
- Katzman S, Kern AD, Bejerano G, Fewell G, Fulton L, Wilson RK, Salama SR, Haussler D. 2007. Human genome ultraconserved elements are ultraselected. *Science* **317**: 915. doi:10.1126/science.1142430
- Kielbasa SM, Wan R, Sato K, Horton P, Frith MC. 2011. Adaptive seeds tame genomic sequence comparison. *Genome Res* **21**: 487–493. doi:10.1101/gr.113985.110
- Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**: 357–360. doi:10.1038/nmeth.3317
- 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. doi:10.1126/science.abn3107
- Kostka D, Hubisz MJ, Siepel A, Pollard KS. 2012. The role of GC-biased gene conversion in shaping the fastest evolving regions of the human genome. *Mol Biol Evol* **29**: 1047–1057. doi:10.1093/molbev/msr279

- Kryuchkova-Mostacci N, Robinson-Rechavi M. 2017. A benchmark of gene expression tissue-specificity metrics. *Brief Bioinformatics* **18**: 205–214.
- Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: a resource for timelines, timetrees, and divergence times. *Mol Biol Evol* **34**: 1812–1819. doi:10.1093/molbev/msx116
- Kvon EZ, Kamneva OK, Melo US, Barozzi I, Osterwalder M, Mannion BJ, Tissières V, Pickle CS, Plajzer-Frick I, Lee EA, et al. 2016. Progressive loss of function in a limb enhancer during snake evolution. *Cell* **167**: 633–642.e11. doi:10.1016/j.cell.2016.09.028
- Lachance J, Tishkoff SA. 2014. Biased gene conversion skews allele frequencies in human populations, increasing the disease burden of recessive alleles. *Am J Hum Genet* **95**: 408–420. doi:10.1016/j.ajhg.2014.09.008
- Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359. doi:10.1038/nmeth.1923
- Lartillot N. 2013. Phylogenetic patterns of GC-biased gene conversion in placental mammals and the evolutionary dynamics of recombination landscapes. *Mol Biol Evol* **30**: 489–502. doi:10.1093/molbev/mst239
- Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 550. doi:10.1186/s13059-014-0550-8
- Mangan RJ, Alsina FC, Mosti F, Sotelo-Fonseca JE, Snellings DA, Au EH, Carvalho J, Sathyan L, Johnson GD, Reddy TE, et al. 2022. Adaptive sequence divergence forged new neurodevelopmental enhancers in humans. *Cell* **185**: 4587–4603.e23. doi:10.1016/j.cell.2022.10.016
- Miller W, Rosenbloom K, Hardison RC, Hou M, Taylor J, Raney B, Burhans R, King DC, Baertsch R, Blankenberg D, et al. 2007. Twenty-eight-way vertebrate alignment and conservation track in the UCSC genome browser. *Genome Res* **17**: 1797–1808. doi:10.1101/gr.6761107
- Mugal CF, Arndt PF, Ellegren H. 2013. Twisted signatures of GC-biased gene conversion embedded in an evolutionary stable karyotype. *Mol Biol Evol* **30**: 1700–1712. doi:10.1093/molbev/mst067
- Necsulea A, Kaessmann H. 2014. Evolutionary dynamics of coding and non-coding transcriptomes. *Nat Rev Genet* **15**: 734–748. doi:10.1038/nrg3802
- 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. doi:10.1093/molbev/msu300
- Pacifici M, Santini L, Di Marco M, Baisero D, Francucci L, Marasini GG, Visconti P, Rondinini C. 2013. Generation length for mammals. *Nat Conserv* **5**: 89–94. doi:10.3897/natureconservation.5.5734
- Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, et al. 2006. In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444**: 499–502. doi:10.1038/nature05295
- Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**: 290–295. doi:10.1038/nbt.3122
- Pessia E, Popa A, Mousset S, Rezvoy C, Duret L, Marais GA. 2012. Evidence for widespread GC-biased gene conversion in eukaryotes. *Genome Biol Evol* **4**: 675–682. doi:10.1093/gbe/evs052
- Poitras L, Yu M, Lesage-Pelletier C, Macdonald RB, Gagné JP, Hatch G, Kelly I, Hamilton SP, Rubenstein JL, Poirier GG, et al. 2010. An SNP in an ultraconserved regulatory element affects *Dlx5/Dlx6* regulation in the forebrain. *Development* **137**: 3089–3097. doi:10.1242/dev.051052
- Pollard KS, Salama SR, King B, Kern AD, Dreszer T, Katzman S, Siepel A, Pedersen JS, Bejerano G, Baertsch R, et al. 2006a. Forces shaping the fastest evolving regions in the human genome. *PLoS Genet* **2**: 1599–1611. doi:10.1371/journal.pgen.0020168
- Pollard KS, Salama SR, Lambert N, Lambot MA, Coppens S, Pedersen JS, Katzman S, King B, Onodera C, Siepel A, et al. 2006b. An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* **443**: 167–172. doi:10.1038/nature05113
- Prabhakar S, Visel A, Akiyama JA, Shoukry M, Lewis KD, Holt A, Plajzer-Frick I, Morrison H, FitzPatrick DR, Afzal V, et al. 2008. Human-specific gain of function in a developmental enhancer. *Science* **321**: 1346–1350. doi:10.1126/science.1159974
- Prum RO, Berv JS, Dornburg A, Field DJ, Townsend JP, Lemmon EM, Lemmon AR. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature* **526**: 569–573. doi:10.1038/nature15697
- Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842. doi:10.1093/bioinformatics/btq033
- Ranwez V, Douzery EJP, Cambon C, Chantret N, Delsuc F. 2018. MACSE v2: toolkit for the alignment of coding sequences accounting for frame-shifts and stop codons. *Mol Biol Evol* **35**: 2582–2584. doi:10.1093/molbev/msy159
- Ratnakumar A, Mousset S, Glémin S, Berglund J, Galtier N, Duret L, Webster MT. 2010. Detecting positive selection within genomes: the problem of biased gene conversion. *Philos Trans R Soc Lond B Biol Sci* **365**: 2571–2580. doi:10.1098/rstb.2010.0007
- Romiguier J, Ranwez V, Douzery EJ, Galtier N. 2010. Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. *Genome Res* **20**: 1001–1009. doi:10.1101/gr.104372.109
- Rousselle M, Laverré A, Figuet E, Nabholz B, Galtier N. 2019. Influence of recombination and GC-biased gene conversion on the adaptive and non-adaptive substitution rate in mammals versus birds. *Mol Biol Evol* **36**: 458–471. doi:10.1093/molbev/msy243
- Sackton TB, Grayson P, Cloutier A, Hu Z, Liu JS, Wheeler NE, Gardner PP, Clarke JA, Baker AJ, Clamp M, et al. 2019. Convergent regulatory evolution and loss of flight in paleognathous birds. *Science* **364**: 74–78. doi:10.1126/science.aat7244
- Seki R, Li C, Fang Q, Hayashi S, Egawa S, Hu J, Xu L, Pan H, Kondo M, Sato T, et al. 2017. Functional roles of Aves class-specific cis-regulatory elements on macroevolution of bird-specific features. *Nat Commun* **8**: 14229. doi:10.1038/ncomms14229
- Snetkova V, Ypsilanti AR, Akiyama JA, Mannion BJ, Plajzer-Frick I, Novak CS, Harrington AN, Pham QT, Kato M, Zhu Y, et al. 2021. Ultraconserved enhancer function does not require perfect sequence conservation. *Nat Genet* **53**: 521–528. doi:10.1038/s41588-021-00812-3
- Snetkova V, Pennacchio LA, Visel A, Dickel DE. 2022. Perfect and imperfect views of ultraconserved sequences. *Nat Rev Genet* **23**: 182–194. doi:10.1038/s41576-021-00424-x
- Stanke J, Moose HE, El-Hodiri HM, Fischer AJ. 2010. Comparative study of Pax2 expression in glial cells in the retina and optic nerve of birds and mammals. *J Comp Neurol* **518**: 2316–2333. doi:10.1002/cne.22335
- Stephen S, Pheasant M, Makunin IV, Mattick JS. 2008. Large-scale appearance of ultraconserved elements in tetrapod genomes and slowdown of the molecular clock. *Mol Biol Evol* **25**: 402–408. doi:10.1093/molbev/msm268
- Storer BE, Kim C. 1990. Exact properties of some exact test statistics for comparing two binomial proportions. *J Am Stat Assoc* **85**: 146–155. doi:10.1080/01621459.1990.10475318
- Sumiyama K, Saitou N. 2011. Loss-of-function mutation in a repressor module of human-specifically activated enhancer HACNS1. *Mol Biol Evol* **28**: 3005–3007. doi:10.1093/molbev/msr231
- Tobias JA, Sheard C, Pigot AL, Devenish AJM, Yang J, Sayol F, Neate-Clegg MHC, Alioravainen N, Weeks TL, Barber RA, et al. 2022. AVONET: morphological, ecological and geographical data for all birds. *Ecol Lett* **25**: 581–597. doi:10.1111/ele.13898
- Torres M, Gómez-Pardo E, Dressler GR, Gruss P. 1995. Pax-2 controls multiple steps of urogenital development. *Development* **121**: 4057–4065. doi:10.1242/dev.121.12.4057
- Torres MA, Gómez-Pardo E, Gruss P. 1996. Pax2 contributes to inner ear patterning and optic nerve trajectory. *Development* **122**: 3381–3391. doi:10.1242/dev.122.11.3381
- Trujillo CA, Rice ES, Schaefer NK, Chaim IA, Wheeler EC, Madrigal AA, Buchanan J, Preissl S, Wang A, Negraes PD, et al. 2021. Reintroduction of the archaic variant of *NOVA1* in cortical organoids alters neurodevelopment. *Science* **371**: eaax2537. doi:10.1126/science.aax2537
- Uebbing S, Gockley J, Reilly SK, Kocher AA, Geller E, Gandotra N, Scharfe C, Cotney J, Noonan JP. 2021. Massively parallel discovery of human-specific substitutions that alter enhancer activity. *Proc Natl Acad Sci* **118**: e2007049118. doi:10.1073/pnas.2007049118
- Ule J, Ule A, Spencer J, Williams A, Hu JS, Cline M, Wang H, Clark T, Fraser C, Ruggiu M, et al. 2005. Nova regulates brain-specific splicing to shape the synapse. *Nat Genet* **37**: 844–852. doi:10.1038/ng1610
- Urbánek P, Fetka I, Meisler MH, Busslinger M. 1997. Cooperation of Pax2 and Pax5 in midbrain and cerebellum development. *Proc Natl Acad Sci* **94**: 5703–5708. doi:10.1073/pnas.94.11.5703
- Vannini I, Wise PM, Challagundla KB, Plousiou M, Raffini M, Bandini E, Fanini F, Paliaga G, Crawford M, Ferracin M, et al. 2017. Transcribed ultraconserved region 339 promotes carcinogenesis by modulating tumor suppressor microRNAs. *Nat Commun* **8**: 1801. doi:10.1038/s41467-017-01562-9
- Visel A, Prabhakar S, Akiyama JA, Shoukry M, Lewis KD, Holt A, Plajzer-Frick I, Afzal V, Rubin EM, Pennacchio LA. 2008. Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat Genet* **40**: 158–160. doi:10.1038/ng.2007.55
- Wang K, Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**: e164. doi:10.1093/nar/gkq603
- Wang Y, Gao S, Zhao Y, Chen W-H, Shao J-J, Wang N-N, Li M, Zhou G-X, Wang L, Shen W-J, et al. 2019a. Allele-specific expression and alternative splicing in horse×donkey and cattle×yak hybrids. *Zoological Research* **40**: 293–304. doi:10.24272/j.issn.2095-8137.2019.042
- Wang Y, Zhang C, Wang N, Li Z, Heller R, Liu R, Zhao Y, Han J, Pan X, Zheng Z, et al. 2019b. Genetic basis of ruminant headgear and rapid antler regeneration. *Science* **364**: eaav6335. doi:10.1126/science.aav6335
- Wang Y, Song C, Zhao J, Zhang Y, Zhao X, Feng C, Zhang G, Zhu J, Wang F, Qian F, et al. 2023. SEdb 2.0: a comprehensive super-enhancer database

- of human and mouse. *Nucleic Acids Res* **51**: D280–D290. doi:10.1093/nar/gkac968
- Weber CC, Boussau B, Romiguier J, Jarvis ED, Ellegren H. 2014. Evidence for GC-biased gene conversion as a driver of between-lineage differences in avian base composition. *Genome Biol* **15**: 549. doi:10.1186/s13059-014-0549-1
- Webster MT, Hurst LD. 2012. Direct and indirect consequences of meiotic recombination: implications for genome evolution. *Trends Genet* **28**: 101–109. doi:10.1016/j.tig.2011.11.002
- Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, Feng T, Zhou L, Tang W, Zhan L, et al. 2021. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation* **2**: 100141.
- Xin Y, Li Z, Zheng H, Ho J, Chan MTV, Wu WKK. 2017. Neuro-oncological ventral antigen 1 (NOVA1): implications in neurological diseases and cancers. *Cell Prolif* **50**: e12348. doi:10.1111/cpr.12348
- Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**: 1586–1591. doi:10.1093/molbev/msm088
- Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S. 2010. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* **26**: 976–978. doi:10.1093/bioinformatics/btq064
- Zhang G, Li C, Li Q, Li B, Larkin DM, Lee C, Storz JF, Antunes A, Greenwold MJ, Meredith RW, et al. 2014. Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* **346**: 1311–1320. doi:10.1126/science.1251385

Received February 8, 2023; accepted in revised form August 23, 2023.