





# Mononucleotide A-repeats may Play a Regulatory Role in Endothermic Housekeeping Genes

Jatuphol Pholtaisong<sup>1</sup> , Nachol Chaiyaratana<sup>2,3</sup> ,  
Chatchawit Aporn Dewan<sup>1,4,5</sup>  and Apiwat Mutirangura<sup>6</sup>

<sup>1</sup>Program in Bioinformatics and Computational Biology, Graduate School, Chulalongkorn University, Pathumwan, Bangkok, Thailand. <sup>2</sup>Department of Electrical and Computer Engineering, Faculty of Engineering, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand. <sup>3</sup>Division of Medical Genetics Research and Laboratory, Research Department, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand. <sup>4</sup>Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University, Pathumwan, Bangkok, Thailand. <sup>5</sup>Omic Sciences and Bioinformatics Center, Chulalongkorn University, Pathumwan, Bangkok, Thailand. <sup>6</sup>Center of Excellence in Molecular Genetics of Cancer and Human Diseases, Department of Anatomy, Faculty of Medicine, Chulalongkorn University, Pathumwan, Bangkok, Thailand.

Evolutionary Bioinformatics  
Volume 18: 1–12  
© The Author(s) 2022  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/11769343221110656  


## ABSTRACT

**BACKGROUND:** Coding and non-coding short tandem repeats (STRs) facilitate a great diversity of phenotypic traits. The imbalance of mononucleotide A-repeats around transcription start sites (TSSs) was found in 3 mammals: *H. sapiens*, *M. musculus*, and *R. norvegicus*.

**PRINCIPAL FINDINGS:** We found that the imbalance pattern originated in some vertebrates. A similar pattern was observed in mammals and birds, but not in amphibians and reptiles. We proposed that the enriched A-repeats upstream of TSSs is a novel hallmark of endotherms or warm-blooded animals. Gene ontology analysis indicates that the primary function of upstream A-repeats involves metabolism, cellular transportation, and sensory perception (smell and chemical stimulus) through housekeeping genes.

**CONCLUSIONS:** Upstream A-repeats may play a regulatory role in the metabolic process of endothermic animals.

**KEYWORDS:** Short tandem repeats, microsatellites, mononucleotide repeats, A-repeats, housekeeping genes, mammals, vertebrates, endotherms

**RECEIVED:** November 22, 2021. **ACCEPTED:** July 2, 2022.

**TYPE:** Original Research

**FUNDING:** The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported in part by the Thailand Research Fund under Grant RSA5980060, RTA6080013, and in part by the Chulalongkorn Academic Advancement into Its 2<sup>nd</sup> Century Project (CUAASC) to CA. The 90<sup>th</sup> Anniversary of Chulalongkorn University Fund (Ratchadaphiseksomphot Endowment Fund) and Science Achievement Scholarship of Thailand (SAST) to JP.

**DECLARATION OF CONFLICTING INTERESTS:** The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**CORRESPONDING AUTHOR:** Chatchawit Aporn Dewan, Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University, Phayathai Road, Pathumwan, Bangkok 10330, Thailand. Email: chatchawit.a@chula.ac.th

## Introduction

A short tandem repeat (STR) is a repetitive sequence of the same unit. The repeat unit is a sequence of nucleotides (A, T, C, or G). For instance, “CAG CAG CAG CAG” is an STR that repeats a trinucleotide unit of “CAG” 4 times. STRs with unit sizes approximately 1 to 6 bp are called microsatellites.<sup>1</sup> The genesis of STRs is due to DNA replication slippage that contracts and expands the length of microsatellites by decreasing and increasing the number of repeat units.<sup>2</sup> Thus, the size of microsatellites is said to be unstable. Microsatellites are found ubiquitously in non-coding regions of eukaryotic genomes. In contrast, long microsatellites in the coding regions are rare, and mutation alters subsequent mRNAs, protein structures, and phenotypes. For instance, Huntington's disease is a well-known phenotype associated with trinucleotide repeats in the exon. The number of CAG units  $\geq 37$  causes a malfunction of Huntingtin protein and associates with a high risk of disease manifestation.<sup>3</sup> In addition, microsatellites may contribute to polygenic diseases<sup>4</sup> and human disorders.<sup>5–9</sup>

Another example is the hexanucleotide repeat that encodes threonine-glycine (Thr-Gly) in the period gene of *Drosophila melanogaster*. The fly's biological clock can adapt to the environmental temperature. For example, the decreased circadian period due to warmer temperatures (18°C–29°C) correlates with the length of Thr-Gly tracts.<sup>10</sup> Similar lines of evidence suggest that coding microsatellites are a source of quantitative genetic variation. Moreover, coding microsatellites may play a crucial role in evolution by equipping genes with adjustable “tuning knobs” for adaptation.<sup>11,12</sup>

On the other hand, non-coding microsatellites were traditionally perceived as non-functional elements.<sup>13</sup> Recent works have increasingly disputed the “selfish DNA” or “junk DNA” dogma by demonstrating that non-coding DNA can mediate the transcription and translation of protein-coding genes.<sup>14</sup> In addition, microsatellites are highly polymorphic.<sup>15,16</sup> The length-variable microsatellites within gene promoters correlated with gene expression variations, which connected with phenotypes.<sup>17</sup> Adaptation via gene modulation is a vital key for



survival in ecological niches and coping with environmental changes. The highly variable microsatellites facilitate a great diversity of phenotypic traits such as cell surface, skeletal morphology, and circadian rhythm.<sup>18,19</sup>

Mononucleotide repeats (unit size=1) are the simplest class, but they are the most abundant microsatellites in genomes. Extensive studies in yeasts showed that non-coding poly(dA:dT) tracts correlate with nucleosome depleted regions.<sup>20,21</sup> These poly(dA:dT) tracts are close to gene promoters and are evolutionarily conserved.<sup>22</sup> Hypothetically, an intrinsic property of poly(dA:dT) tracts is to resist sharp DNA bending into the helical structure. The crystal structure of “CGCAAAAAGCG” showed that the poly(dA:dT) tract is essentially straight and distinctive from other DNA sequences.<sup>23</sup> Runs of about 5 bp of the homopolymer phased every 10 to 11 bp (every helical turn) alter the DNA configuration required for some protein-binding activities. For instance, SV40 large T antigen requires 2 pentanucleotide repeats (5'-GAGGC-3'/5'-GCCTC-3') and the asymmetric sequence (5'-TTTTTTG-3'/5'-CAAAAAA-3') that separates the pentanucleotide repeats.<sup>24</sup>

About 85% of 696 016 microsatellites in the human genome are conserved in at least one other species (11 mammalian and 5 non-mammalian vertebrates).<sup>25</sup> In addition, the imbalance of sense mononucleotide A-repeats around transcription start sites (TSSs) was observed.<sup>26</sup> A comparison between upstream and downstream of TSSs showed that the number of long A-repeats ( $\geq 10$ bp) was disproportionate in 3 mammals (*Rattus norvegicus*, *Mus musculus*, *Homo sapiens*), but not in 3 non-mammals (*Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*). Thus, the imbalance pattern must originate somewhere before the rise of mammals and primates. The previous work suggests that the pattern should be found in higher animals but not in lower organisms such as yeasts, worms, and flies. Investigating the genesis of this imbalance pattern will reveal the biological function of A-repeats. We explored the genomes of 60 organisms in the National Center for Biotechnology Information (NCBI) database<sup>27</sup> (Figure 1). Finally, we could identify A-repeat enriched genes and speculate the physiological function of A-repeats.

## Materials and Methods

### Genomic data

Initially, the reference genomes of all organisms in the NCBI database<sup>27</sup> were considered. Next, we filtered only the organisms that were adequately annotated (at least 10 000 genes with RefSeq status “model” or “reviewed” or “validated” or “provisional”). Then, we selected vertebrate genomes with the highest number of genes in each group (mammals, birds, reptiles, amphibians, and fish) but no more than 20 genomes per group. As a result, 20 mammals, 13 birds, 3 reptiles, 4 amphibians, and 20 fish were included in our analysis (Supplemental Table S1). An in-house computer program was developed to count

perfect mononucleotide repeats (A-, T-, C-, and G-repeats) in the genomes. Since perfect mononucleotide repeats are easy to count, we used a brute-force algorithm that runs in a linear time with genome size. Unlike imperfect and higher-order repeats, no advanced data structures or computational techniques were needed.

### Housekeeping and tissue-specific genes

A total of 2108 human housekeeping genes were retrieved from the Housekeeping and Reference Transcript Atlas (HRT 1.0).<sup>28</sup> The list of 2833 tissue-specific genes in humans was compiled from a map of the human tissue proteome.<sup>29</sup>

### Binning method

The genomic sequences in 5000bp upstream and 5000bp downstream of the transcription start sites (TSSs) were divided into 20 bins (Supplemental Figure S1). The selected bin size (500bp) was suitable for studying long repeats ( $\geq 10$ bp), which did not occur very often. Subsequently, bins 1 to 10 were the upstream region, and bins 11 to 20 were the downstream region. Only repeats on the sense strands were counted; otherwise the imbalance pattern could not be observed. A repeat that spanned 2 bins was counted in the base pair (bp) unit and proportionally accumulated in each bin (Supplemental Figure S2). The number of repeats was normalized by dividing by the total number of genes. As a result, the number of repeats in each bin was measured in the unit of base pairs per gene. In the final step, each bin was divided by the total number of genes so that the number of repeats between 2 unequal sets of genes could be compared. Genes could overlap to each other. In 2009, there were not many overlapping genes. In humans, there were 774 pairs of overlapping genes among 34 604 annotated genes.<sup>30</sup> However, the current number of overlapping genes increases drastically. Therefore, we decided to remove all pairs of overlapping genes. The numbers of non-overlapping genes are provided in Supplemental Table S1.

### Nearest TSS method

The nearest TSS method took a single repeat as input and searched for the nearest TSS in the 3' direction of the repeat. Note that the repeat and the nearest TSS (gene) must be on the same strand so that the repeat was upstream of TSS. The distance between a repeat and the nearest TSS was measured in base pairs (bp).

### Gene ontology (GO) analysis

We conducted GO analysis of 3 human gene sets: 2108 housekeeping genes, 2833 tissue-specific genes, and the top 10% upstream A-repeat enriched genes (1871 genes). We counted only the upstream repeats in the first to the seventh bins

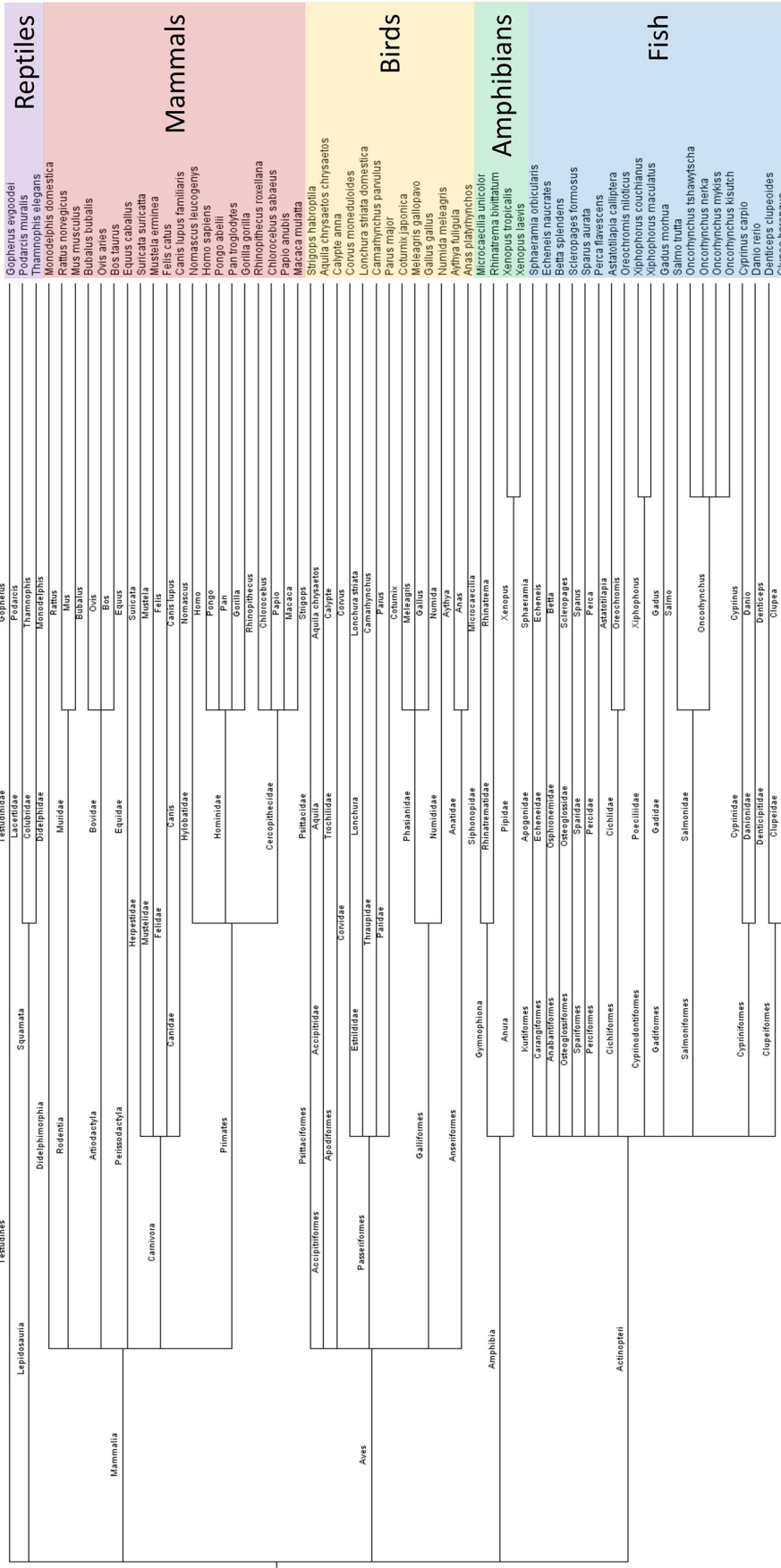


Figure 1. The phylogenetic tree of 60 selected vertebrate species.

(3500bp), where A-repeats were maximal. Each gene set was submitted to the PANTHER GO Enrichment Analysis with default parameters (Fisher's exact test and false discovery rate).<sup>31</sup> The "GO biological process complete" was chosen for annotation data set. We also conducted GO analysis in other 13 species (*Pan troglodytes*, *Macaca mulatta*, *Gorilla gorilla*, *Mus musculus*, *Rattus norvegicus*, *Monodelphis domestica*, *Felis catus*, *Canis lupus familiaris*, *Equus caballus*, *Bos taurus*, *Gallus gallus*, *Xenopus laevis*, and *Danio rerio*) that were available in the PANTHER GO Enrichment Analysis.<sup>31</sup> The resulting GO terms are listed in Supplemental Table S2.

### Phylogenetic tree

The phylogenetic tree in Figure 1 was built using the NCBI Taxonomy.<sup>32</sup> Thereafter, the phylogenetic tree file (.phy) was visualized using the FigTree software (<http://tree.bio.ed.ac.uk/software/figtree>).

### Statistical analysis

The Kruskal-Wallis test, the independent *t*-test, the Mann-Whitney *U* test, the Spearman's correlation, and the Fisher's exact test were conducted using SPSS version 28.<sup>33</sup> Note that the Mann-Whitney *U* test was used instead of the independent *t*-test when the dependent variable was not normally distributed.

## Results

### Distribution of repeat length in mammals, birds, reptiles, amphibians, and fish

Mononucleotide repeats were ubiquitous in the vertebrate genomes (Figure 2). The distributions of repeats in mammals, birds, reptiles, amphibians, and fish are not identical. The Kruskal-Wallis test *P*-values are .006, .003, .007, and .004, respectively for A-, T-, C-, and G-repeats. Long A- and T-repeats were the highest in mammals and birds (Figure 2A and B), and declined in fish, amphibians, and reptiles, respectively. On the other hand, long C- and G-repeats were the most abundant in amphibians and fish (Figure 2C and D), and declined in birds, mammals, and reptiles, respectively. All classes of mononucleotide repeats were the lowest in reptiles. The numbers of short repeats did not differ among the vertebrate groups, while the numbers of long repeats were distinguishable. The threshold for long A- and T-repeats was  $\geq 10$ bp, which was concordant with the reported findings for the biological function of A-repeats.<sup>26</sup> For short repeats, the number of repeats exponentially decreased with increasing repeat length, as indicated by a straight line in the logarithmic scale. However, a sudden surge of the slope was observed in all graphs (Figure 2). The excessive number of long A- and T-repeats suggests that the repeats may play an essential role in the avian and mammalian genomes. The highest number of

C- and G-repeats found in amphibians and fish was unexpected because mononucleotide repeats were more frequent in humans and primates.<sup>34</sup> The previous survey<sup>34</sup> of microsatellites in different eukaryotic genomes did not report the number of C- and G-repeats in amphibians and fish because amphibians and fish were grouped to vertebrata. In addition, A-, T-, C-, and G-repeats were grouped to mononucleotide repeats. Interestingly, long C- and G-repeats were abundant around TSSs in reptiles (Figure 3E), and gradually depleted in amphibians, fish, and birds, respectively (Figure 3D, F, and G). The number of long C- and G-repeats significantly dropped in mammals, and was the lowest in primates. These findings are worthy for further investigation about the role of C- and G-repeats before the rise of mammals and primates. However, this paper will focus on A-repeats in vertebrates, specifically warm-blooded animals.

### Distribution of long repeats around TSSs

We divided the 10000bp around TSSs into 20 bins and counted the number of long repeats ( $\geq 10$ bp) in each bin (see Materials and Methods). We found that the number of A- and T-repeats dropped very sharply around TSSs in all mammals (Figure 3A-C). The mean differences between the upstream (-5000 to -1500bp) and downstream (1500-5000bp) long-A repeats are 717.88 (mammals), 985.45 (primates), and 450.31 (non-primate mammals). The corresponding Mann-Whitney *U* test *P*-values are all  $5.83e-04$ . The mean differences of T-repeats are -204.33 (mammals), -31.66 (primates), and -377.01 (non-primate mammals). The corresponding *t*-test *P*-values are  $1.19e-03$ ,  $6.80e-01$ , and  $5.87e-13$ , respectively. In the vertebrate genomes, CpG islands were associated with the 5' ends of all housekeeping genes and many tissue-specific genes.<sup>35,36</sup> Therefore, the CpG islands could disrupt the occurrence of long A- and T-repeats. The sharp drop of repeats below 1000bp/Mbp and the imbalance between upstream and downstream A-repeats were evident in mammals (Figure 3A-C) but less apparent in other non-mammalian vertebrates (Figure 3D-G). The mean differences between the upstream (-5000 to -1500bp) and downstream (1500-5000bp) long-A repeats are 378.98 (birds), 16.71 (reptiles), -165.46 (amphibians), and 165.08 (fish). The corresponding *t*-test *P*-values are  $5.58e-13$ ,  $1.53e-01$ ,  $5.50e-05$ , and  $1.33e-11$ , respectively. The mean differences of T-repeats are -464.75 (birds), -120.21 (reptiles), -702.97 (amphibians), and -243.83 (fish). The corresponding *t*-test *P*-values are  $2.57e-10$ ,  $8.59e-09$ ,  $8.44e-13$ , and  $2.44e-08$ , respectively.

### Distribution of long A-repeats in exon, intron, CDS, 5' UTR, and 3' UTR of genes

Long A-repeats almost disappeared from the coding sequence (CDS) (Figure 4). In addition, the number of repeats in exons was low because exons were composed of CDSs and

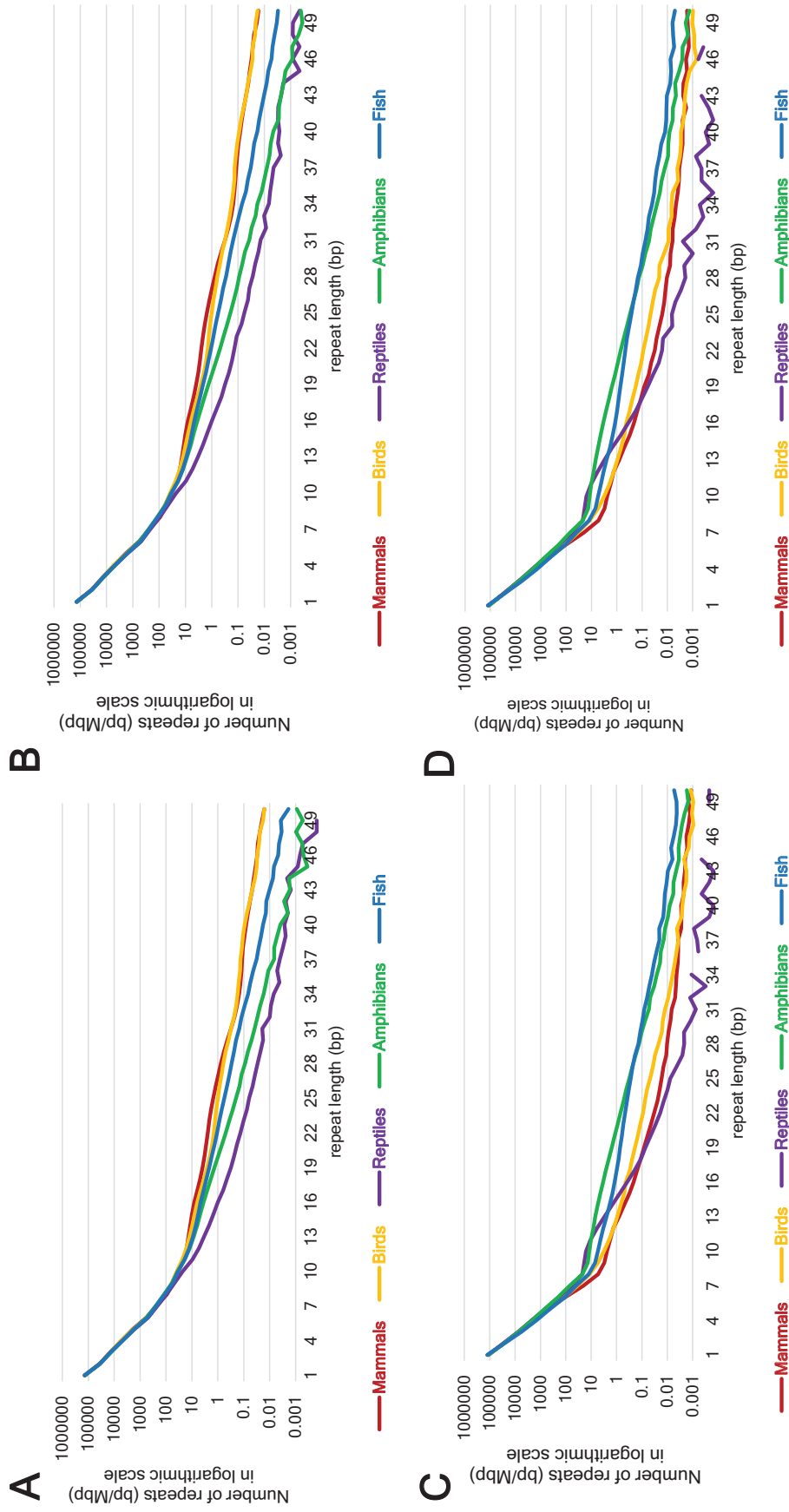
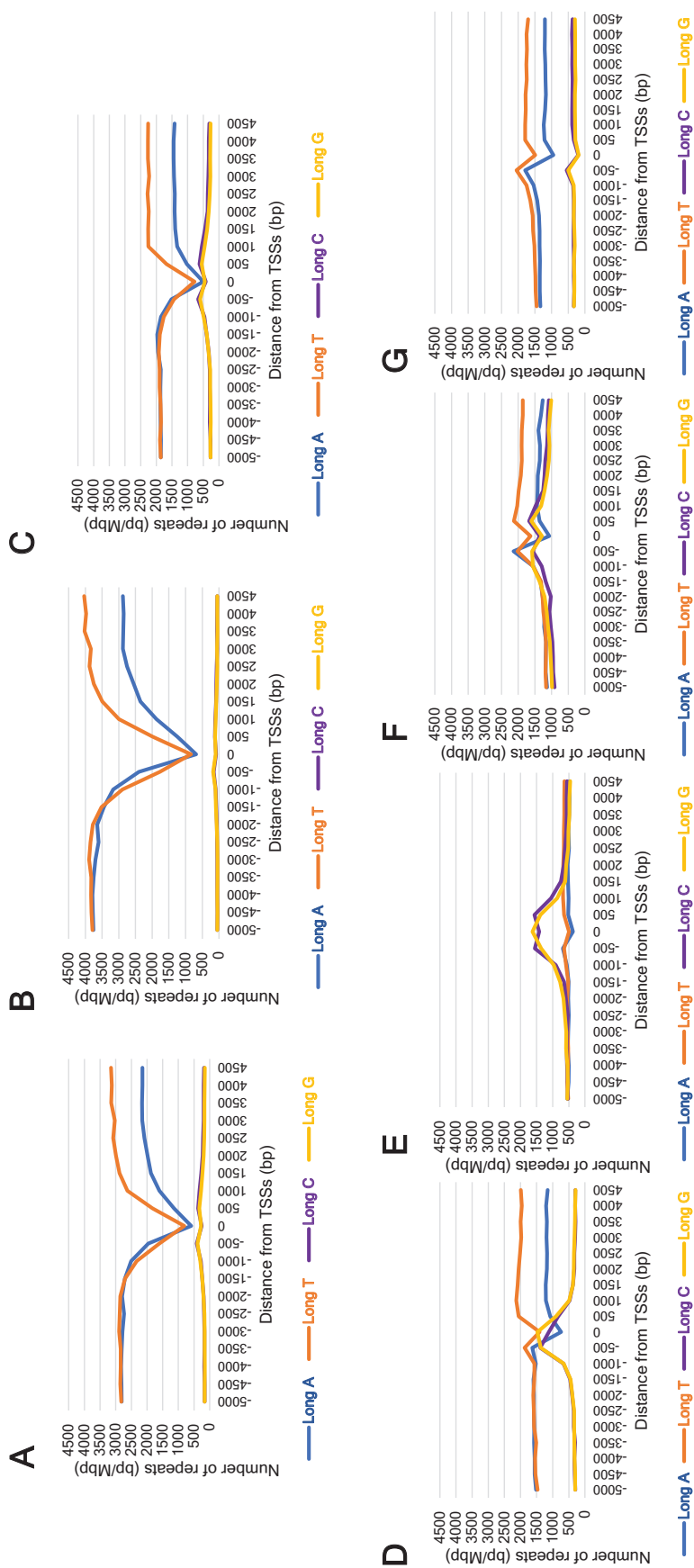
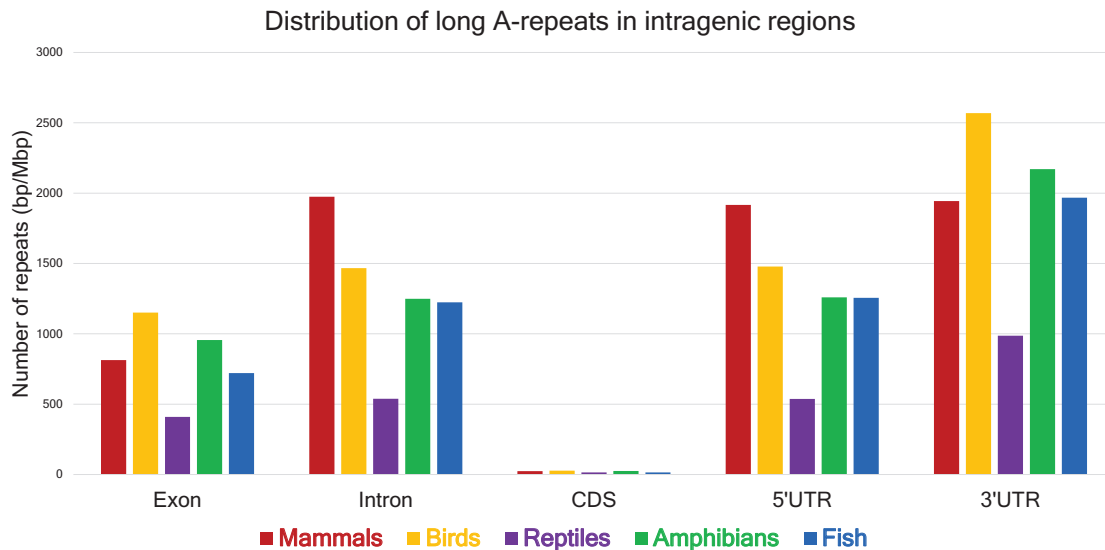


Figure 2. Distribution of repeat length in mammals, birds, reptiles, amphibians, and fish: (A) A-repeats, (B) T-repeats, (C) C-repeats, and (D) G-repeats.





**Figure 3.** Distribution of long repeats ( $\geq 10$  bp) around transcription start sites (TSSs) in mammals, birds, reptiles, amphibians, and fish: (A) mammals, (B) non-primate mammals, (C) birds, (D) reptiles, (E) amphibians, and (G) fish.



**Figure 4.** The number of long A-repeats ( $\geq 10$ bp) in 5'UTR, exon, CDS, intron, and 3'UTR in mammals, birds, reptiles, amphibians, and fish.

untranslated regions. Introns, 5'UTR, and 3'UTR showed a similar pattern. Intragenic A-repeats were the highest in mammals and birds. The number of A-repeats in amphibians and fish were almost identical, whereas approximately 50% reduction in reptiles.

#### *Distribution of long A-repeats in human housekeeping and tissue-specific genes*

The human housekeeping genes showed a distinctive distribution (Figure 5). First, long A- and T-repeats were more abundant in the housekeeping genes ( $\sim 2$  times compared with tissue-specific genes). Second, long A-repeats were enriched in the upstream of housekeeping genes. The mean differences between the upstream ( $-5000$  to  $-1500$ bp) and downstream ( $1500$ - $5000$ bp) long-A repeats are 1572.43 (house-keeping genes) and 675.87 (tissue-specific genes). The corresponding  $t$ -test  $P$ -values are  $7.11e-06$  and  $1.02e-03$ , respectively. The mean differences of T-repeats are  $-446.81$  (house-keeping genes) and 382.60 (tissue-specific genes). The corresponding  $t$ -test  $P$ -values are  $2.62e-02$  and  $8.03e-03$ , respectively. In Figure 5A, it is not clear whether A-repeats were enriched upstream of TSSs, or A-repeats were suppressed downstream. Thus, we conducted the experiment in Figure 6 specifically to answer this question.

#### *The enrichment of upstream A-repeats in human housekeeping genes*

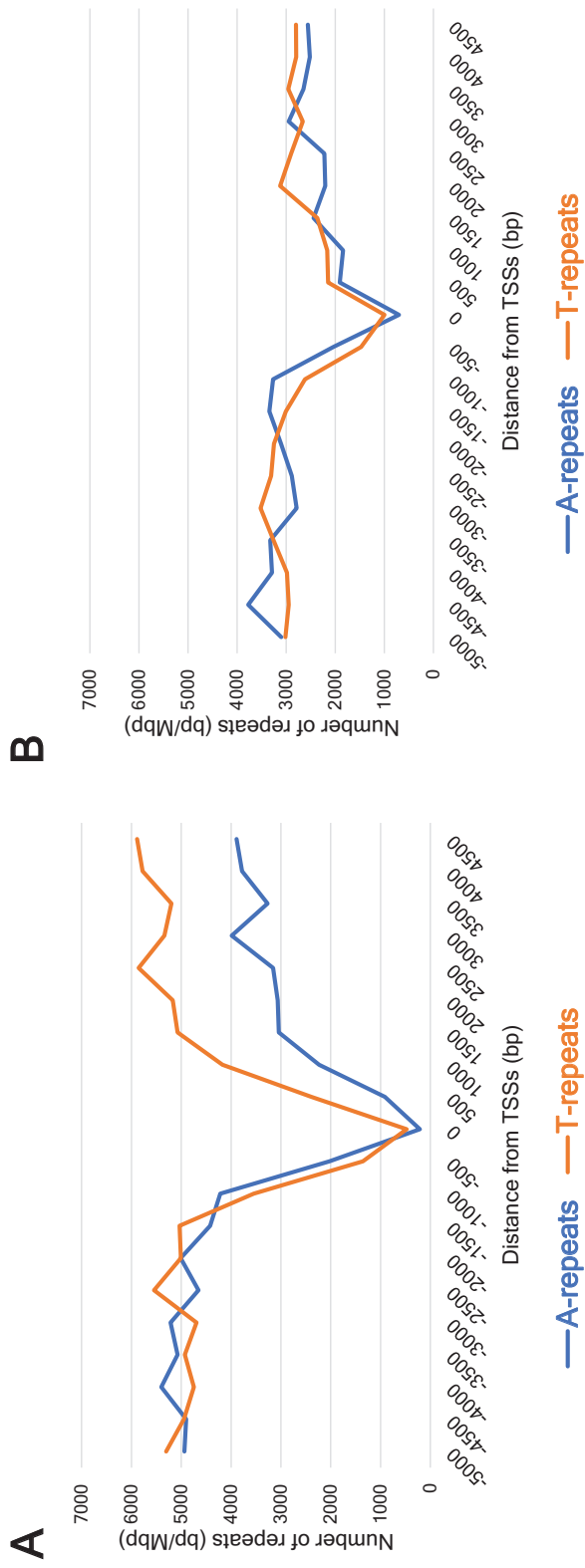
The upstream A- and T-repeat locations were not random, but the occurrence of repeats increased with the proximity to the TSSs of human housekeeping genes (Figure 6A). The x-axis is the distance between a repeat to the nearest TSSs, and the y-axis is the observed frequency of repeats (see Materials and Methods). The correlation from  $-50000$  bp to  $-2500$  bp in housekeeping genes is moderate, Spearman's  $r_s$ :

$-0.517$  (A-repeats) and  $-0.645$  (T-repeats) with corresponding  $P$ -values:  $3.62e-08$  and  $4.42e-13$ , respectively. The low frequencies in the first few bins indicate that long repeats rarely occurred close to the TSSs. The A- and T-repeat frequencies went to the highest peak at the  $\sim 2500$ bp upstream of TSSs. In the more upstream regions, the occurrence of repeats declined with the distance from TSSs. In contrast, a similar pattern was not found in the tissue-specific genes (Figure 6B). The correlation from  $-50000$  bp to  $-2500$  bp in tissue-specific genes is weak, Spearman's  $r_s$ :  $-0.250$  (A-repeats) and  $-0.363$  (T-repeats) with corresponding  $P$ -values:  $1.21e-02$  and  $2.02e-04$ , respectively.

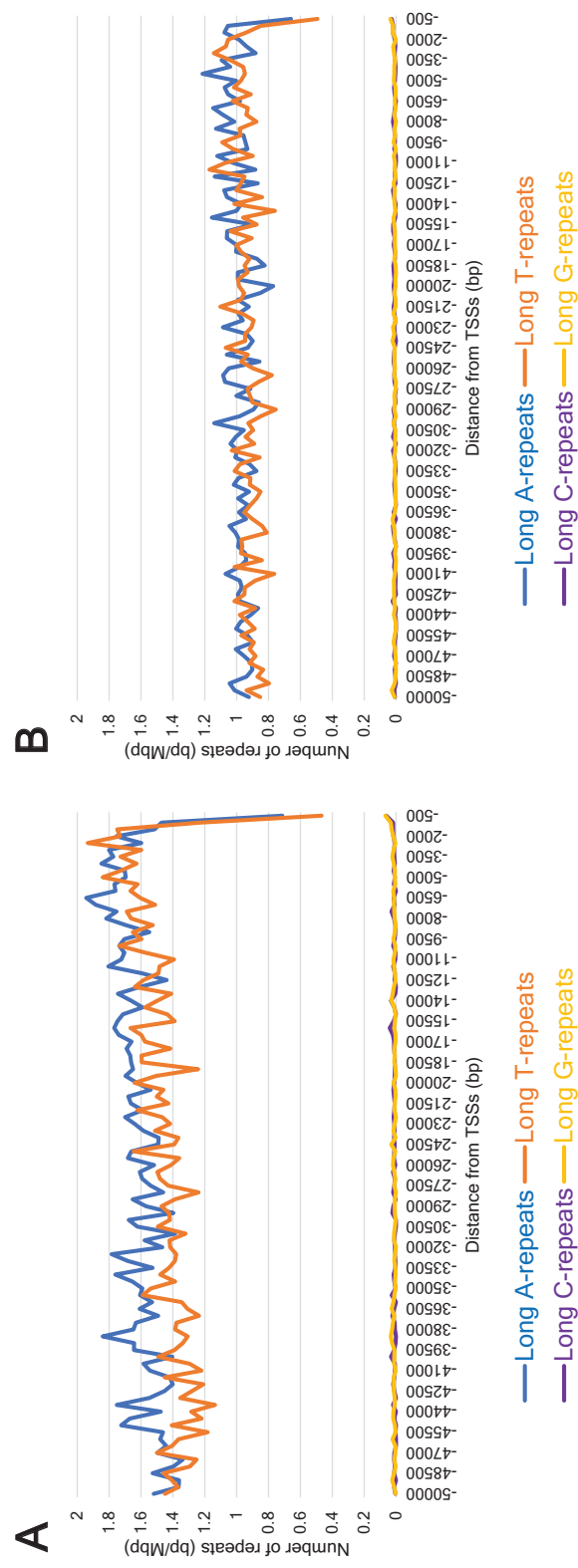
In human housekeeping genes, it is clear that the number of long A-repeats is higher than the number of T-repeats (Figure 6A). The mean difference is 0.15 and the corresponding  $t$ -test  $P$ -value is  $2.16e-09$ . In contrast, tissue-specific genes show less mean difference, 0.05, and the corresponding  $t$ -test  $P$ -value is  $1.29e-04$ . Figure 6 confirms that A- and T-repeats were enriched upstream of human housekeeping genes and A-repeats were more enriched than T-repeats.

#### *GO enrichment analysis*

Human housekeeping genes were found in a broad range of GO terms, for instance, GO:0006412 translation (29.1%), GO:0003735 structural constituent of ribosome (23.3%), and GO:0005840 ribosome (18.4%).<sup>37</sup> To identify the specific GO terms, we conducted the GO enrichment analysis (see Materials and Methods) for 3 gene sets in humans: housekeeping (HK) genes, tissue-specific (TS) genes, and the top 10% upstream A-repeat enriched (top-A) genes (Figure 7). Housekeeping genes largely overlapped with top-A genes (Figure 7A). The corresponding  $2 \times 2$  contingency table of {HK, TS}  $\times$  {Top-A, not Top-A} yields  $a$ : 250 + 1,  $b$ : 1736 + 3,  $c$ : 110 + 1,  $d$ : 1738 + 3, OR: 2.26, 95% CI: 1.78 to 2.88, and Fisher's exact test  $P$ -value:  $1.42e-12$ . As a result, housekeeping genes had

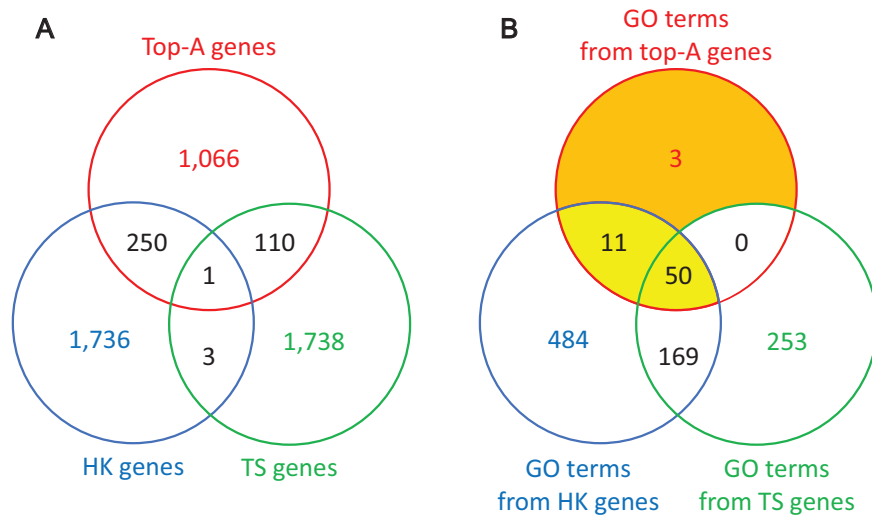


**Figure 5.** Distribution of long A- and T-repeats ( $\geq 10$  bp) around transcription start sites of human housekeeping and tissue-specific genes: (A) housekeeping genes and (B) tissue-specific genes.



**Figure 6.** The occurrence of long repeats ( $\geq 10$  bp) varying with the distance to the nearest transcription start sites of human housekeeping and tissue-specific genes: (A) housekeeping genes and (B) tissue-specific genes.





**Figure 7.** (A) The numbers of genes and (B) GO terms from housekeeping (HK) genes, tissue-specific (TS) genes, and the top 10% upstream A-repeat enriched (top-A) genes in humans.

more than two times the odds of being top-A genes, compared with tissue-specific genes. Next, we intersected the GO terms from HK, TS, and top-A genes (Figure 7B). Most GO terms from top-A genes intersect with those from HK genes ( $(50 + 11) \div (50 + 11 + 3) = 95.3\%$ ), and less intersect with those from TS genes ( $(50 \div (50 + 11 + 3) = 78.1\%$ ). Many GO terms from TS genes involved because a biological process may require both housekeeping and tissue-specific genes. The overlap between GO terms from HK and TS genes was a proof. We concluded that upstream A-repeats primarily play a regulatory role through housekeeping genes. Although some TS genes are top-A genes, all GO terms from TS genes that overlap with the GO terms from top-A genes absolutely overlap with the GO terms from HK genes. The  $50 + 11 = 61$  biological processes (yellow color in Figure 7B) cover metabolic processes, cellular transportation, and detection of stimulus involved in sensory perception, etc. (Supplemental Table S2). The three GO terms (orange color in Figure 7B) that do not overlap with those from HK genes are chromosome segregation, neuron projection guidance, and axon guidance (Supplemental Table S2).

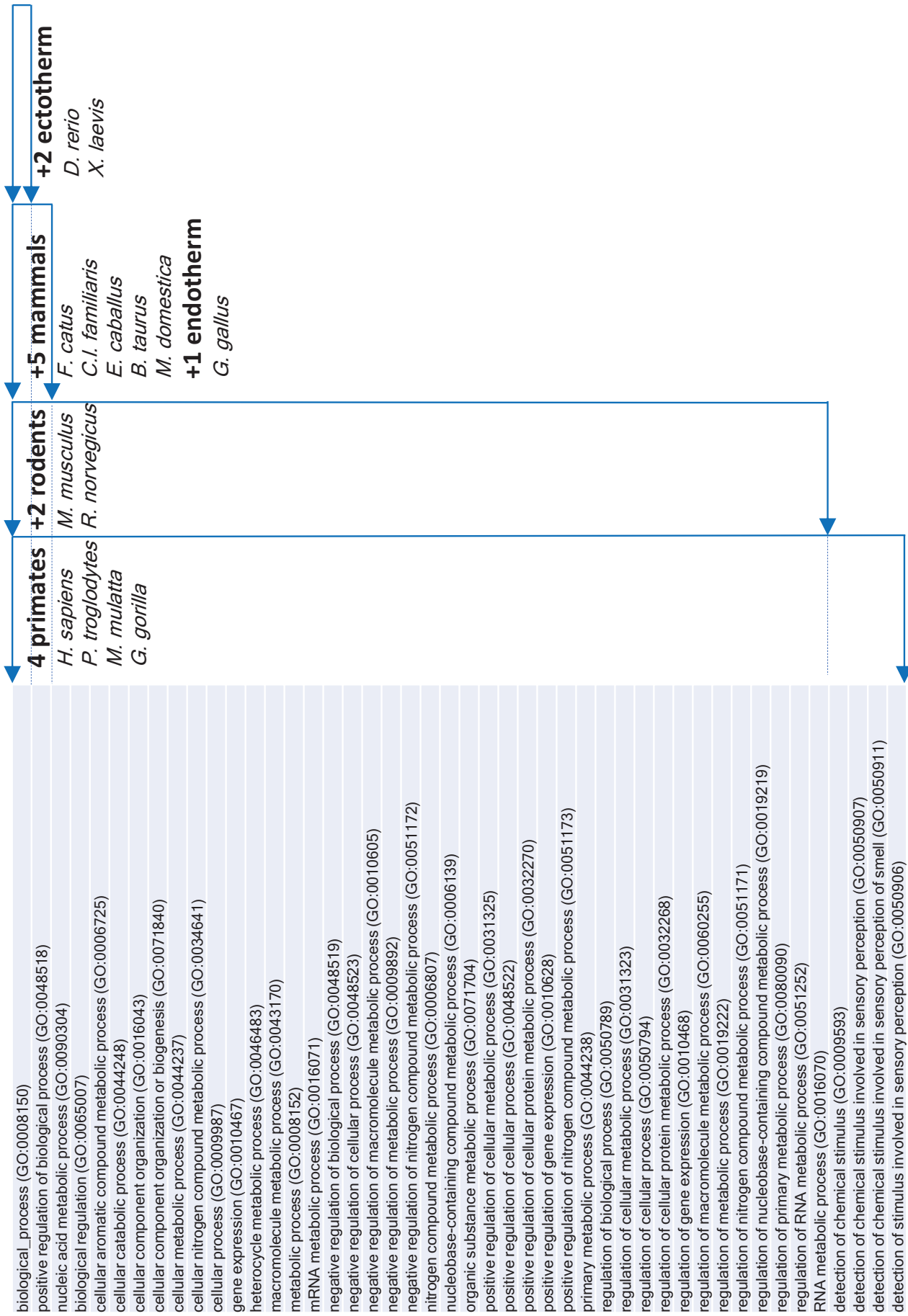
We extended the GO analysis to other 13 species that were available in the PANTHER GO Enrichment Analysis.<sup>31</sup> The homologous genes between humans and other species are provided in Supplemental Table S3. *Pan troglodytes*, *Mus musculus*, *Rattus norvegicus*, *Canis lupus familiaris*, *Bos taurus*, *Macaca mulatta*, *Xenopus tropicalis*, *Gallus gallus*, and *Danio rerio* have 17359 (92.8%), 16766 (89.6%), 16289 (87.0%), 16288 (87.0%), 16198 (86.6%), 15297 (81.7%), 12500 (66.8%), 12220 (65.3%), and 12107 (64.2%) homologs, respectively. Figure 8 shows the intersection of GO terms from a number of taxonomic groups. The GO terms were derived solely from the top-A genes. The threshold of top-A genes was set at 10% as in humans. The positive regulation of biological process (GO:0048518) is the first common GO term in endotherms. The role of upstream A-repeats in metabolic processes begins

in mammals and rodents. Subsequently, the role of detection of stimulus involved in sensory perception appears in primates.

## Discussion

The previous work discovered the imbalance between upstream and downstream A-repeats in humans and suggested a length-dependent *cis*-regulatory function of A-repeats, with Ago proteins as *trans*-acting factors.<sup>26</sup> Herein, we found that the imbalance pattern is a common characteristic of 20 mammalian genomes (Figure 3A–C). Furthermore, a similar pattern in 13 birds and 20 fish (Figure 3D and G) suggests that the imbalance of upstream and downstream A-repeats originated in some non-mammalian vertebrates. On the other hand, we did not observe such a pattern in 3 reptiles and 4 amphibians (Figure 3E and F). A cluster of CpG islands at TSSs is a hallmark of housekeeping genes in vertebrates.<sup>35,36</sup> CpG islands interplay with mononucleotide repeats by disrupting each other. Nevertheless, we found that the enrichment of upstream A-repeats is a novel hallmark of endotherms or warm-blooded animals (mammals and birds) and fish despite the underlying biological functions and mechanisms.

Fish are the first vertebrate that originated during the Cambrian explosion 541 million years ago (mya). The transition from fish to amphibians occurred in the Devonian period (419.2 mya). Next, reptiles emerged in the Carboniferous period (358.9 mya). The evolution of endothermy began ~250 mya from 2 different groups of reptilian ancestors: the sauropsid lineage (birds) and the synapsid lineage (mammals).<sup>38</sup> In general, fish are ectothermic or cold-blooded, but some can produce internal heat during embryo development,<sup>39</sup> and some fast-swimming fish are endothermic.<sup>40</sup> The upstream A-repeats (Figure 3A, D, and G) are concordant with the evolution of endothermy. The pattern slightly differs between mammals and birds, which independently descend from the 2 lineages. The degraded pattern in fish is consistent with the fact that



**Figure 8.** The intersection of GO terms from top 10% upstream A-repeat enriched (top-A) genes in the species that are available in the PANTHER GO Enrichment Analysis. The first intersection includes 4 primates: *Homo sapiens*, *Pan troglodytes*, *Macaca mulatta*, and *Gorilla gorilla*. The next intersections increase the number of species by adding 2 rodents (*Mus musculus* and *Rattus norvegicus*), 5 mammals (*Felis catus*, *Canis lupus familiaris*, *Equus caballus*, *Bos taurus*, and *Monodelphis domestica*), 1 endotherm (*Gallus gallus*), and 2 ectotherms (*Danio rerio* and *Xenopus laevis*), respectively.

fish are endothermic under some circumstances. Although amphibians and reptiles branched from fish, their pattern of upstream A-repeats is not similar to that of fish. Endothermy necessitates tremendous regulation of housekeeping genes to satisfy the metabolic requirements in various tissues. Our findings suggest that upstream A-repeats are a regulatory element for endothermic housekeeping genes. More precisely, the primary target of upstream A-repeats is not only housekeeping genes, but the housekeeping processes. The metabolism, cellular transportation, and sensory perception (smell and chemical stimulus) seem to be the target of A-repeats, as indicated by significant GO terms (Supplemental Table S2). The olfactory sense is linked to metabolism. For example, the sense of smell affects aging in worms and flies.<sup>41</sup> In addition, mice lacking olfactory sensory neurons are resistant to diet-induced obesity.<sup>42</sup> The secondary targets of A-repeats are chromosome segregation, neuron projection guidance, and axon guidance as indicated by the GO terms that do not overlap with the housekeeping processes. These GO terms can be summarized as the development of nervous system, and significantly enriched with upstream A-repeats only in humans.

If the “junk DNA” dogma is confirmed, A-repeats possess no functions. In this point of view, the existence of A-repeats is to replace other regulatory elements that suppress gene expression. Since housekeeping genes require a constant expression level, a large number of repeats might be recruited to fill in gene promoters and the upstream. However, there are several arguments. First, if the repeats are junk, either A- or T-repeats are the same, but upstream A-repeats are more frequent than T-repeats. Perhaps there is a constraint on genomic architecture that inhibits upstream T-repeats. Second, poly(dA:dT) tracts alter DNA's helical structure.<sup>23</sup> Recent research showed that the three-dimensional structure of a DNA-binding site determines the shape of transcription factor binding, which in turn influences the transcription activity.<sup>43</sup> Third, A-repeats are *cis*-regulatory elements, and Argonaute proteins serve as *trans*-acting factors.<sup>26</sup> Argonaute proteins (AGO1, AGO2, AGO3, AGO4) preferentially bind A-repeats. Moreover, the binding preference increases with repeat length. Targeted therapy via A-repeats has enormous potential for regulating housekeeping genes in concert. For example, transfection of molecules that mimic A-repeats to lung cancer cells and lung cancer stem cells inhibited cell proliferation and prevented a single cell from growing to a colony.<sup>44,45</sup>


In conclusion, we discovered the origin of the imbalance between upstream and downstream A-repeats. A similar pattern found in mammals, birds, and fish but not in reptiles and amphibians suggests that upstream A-repeats are a characteristic of endothermic housekeeping genes. In humans, housekeeping genes are the main contributor to upstream A-repeats. Unfortunately, we could not find a complete list of housekeeping and tissue-specific genes for all vertebrates. The human reference genome is the most reliable because it has been sequenced across millions of individuals over the recent decades. The genome quality of laboratory animals

and livestock are next to that of humans. On the other hand, the reference genomes of some species are still far from completed and fully annotated. The future progress of genome annotation and gene expression databases is needed to confirm our findings. As pointed out by an anonymous reviewer, the variance in these mononucleotide repeats should be compared in publicly available human genome databases. The extent of polymorphism and mutation across diverse human genomes globally is of great interest. Furthermore, it could help address their hypothesis regarding roles in gene classes involved in endotherms or other biological processes.

### Author Contributions

Conceived and designed the experiments: JP, CA, AM. Collected and analyzed the data: JP. Contributed to the writing of the manuscript: CA, NC. All authors reviewed and approved of the final manuscript.

### ORCID iDs

Jatuphol Pholtaisong  <https://orcid.org/0000-0002-2575-8366>

Nachol Chaiyaratana  <https://orcid.org/0000-0001-5194-8801>

Chatchawit Aporntewan  <https://orcid.org/0000-0001-7290-8286>

### Supplemental Material

Supplemental material for this article is available online.

### REFERENCES

- Ellegren H. Microsatellites: simple sequences with complex evolution. *Nat Rev Genet.* 2004;5:435-445.
- Schlötterer C, Tautz D. Slippage synthesis of simple sequence DNA. *Nucleic Acids Res.* 1992;20:211-215.
- DiFiglia M, Sapp E, Chase KO, et al. Aggregation of huntingtin in neuronal intranuclear inclusions and dystrophic neurites in brain. *Science.* 1997;277:1990-1993.
- Hannan AJ. Tandem repeat polymorphisms: modulators of disease susceptibility and candidates for ‘missing heritability’. *Trends Genet.* 2010;26:59-65.
- Hannan AJ. Repeat DNA expands our understanding of autism spectrum disorder. *Nature.* 2021;589:200-202.
- Herbert A. Simple repeats as building blocks for genetic computers. *Trends Genet.* 2020;36:739-750.
- Rodriguez CM, Todd PK. New pathologic mechanisms in nucleotide repeat expansion disorders. *Neurobiol Dis.* 2019;130:104515.
- Gonzalez-Alegre P. Recent advances in molecular therapies for neurological disease: triplet repeat disorders. *Hum Mol Genet.* 2019;28:R80-R87.
- Hannan AJ. Tandem repeats mediating genetic plasticity in health and disease. *Nat Rev Genet.* 2018;19:286-298.
- Sawyer LA, Hennessy JM, Peixoto AA, et al. Natural variation in a *Drosophila* clock gene and temperature compensation. *Science.* 1997;278:2117-2120.
- King DG, Soller M, Kashi Y. Evolutionary tuning knobs. *Endeavour.* 1997;21:36-40.
- King DG, Soller M. Variation and fidelity: the evolution of simple sequence repeats as functional elements in adjustable genes. In: Wasser SP, ed. *Evolutionary Theory and Processes: Modern Perspectives: Papers in Honour of Eviatar Nevo.* Springer; 1999:65-82.
- Hall SS. Journey to the genetic interior. *Sci Am.* 2012;307:80-84.
- ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature.* 2012;489:57-74.
- Duitama J, Zablotskaya A, Gemayel R, et al. Large-scale analysis of tandem repeat variability in the human genome. *Nucleic Acids Res.* 2014;42:5728-5741.

16. Legendre M, Pochet N, Pak T, Verstrepen KJ. Sequence-based estimation of minisatellite and microsatellite repeat variability. *Genome Res.* 2007;17:1787-1796.
17. Vences MD, Legendre M, Caldara M, Hagihara M, Verstrepen KJ. Unstable tandem repeats in promoters confer transcriptional evolvability. *Science.* 2009;324:1213-1216.
18. Gemayel R, Vences MD, Legendre M, Verstrepen KJ. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu Rev Genet.* 2010;44:445-477.
19. Gemayel R, Cho J, Boeynaems S, Verstrepen KJ. Beyond junk-variable tandem repeats as facilitators of rapid evolution of regulatory and coding sequences. *Genes.* 2012;3:461-480.
20. Segal E, Widom J. Poly(dA:dT) tracts: major determinants of nucleosome organization. *Curr Opin Struct Biol.* 2009;19:65-71.
21. Rando OJ, Winston F. Chromatin and transcription in yeast. *Genetics.* 2012;190:351-387.
22. Yuan GC, Liu YJ, Dion MF, et al. Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science.* 2005;309:626-630.
23. Nelson HC, Finch JT, Luisi BF, Klug A. The structure of an oligo(dA).oligo(dT) tract and its biological implications. *Nature.* 1987;330:221-226.
24. Ryder K, Silver S, DeLucia AL, Fanning E, Tegtmeyer P. An altered DNA conformation in origin region I is a determinant for the binding of SV40 large T antigen. *Cell.* 1986;44:719-725.
25. Buschiazzo E, Gemmell NJ. Conservation of human microsatellites across 450 million years of evolution. *Genome Biol Evol.* 2010;2:153-165.
26. Aporntewan C, Pin-on P, Chaiyaratana N, Pongpanich M, Boonyaratana-kornkit V, Mutirangura A. Upstream mononucleotide A-repeats play a cis-regulatory role in mammals through the DICER1 and Ago proteins. *Nucleic Acids Res.* 2013;41:8872-8885.
27. Sayers EW, Beck J, Bolton EE, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 2021;49:D10-D17.
28. Hounkpe BW, Chenou F, de Lima F, De Paula EV. HRT Atlas v1.0 database: redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive RNA-seq datasets. *Nucleic Acids Res.* 2021;49:D947-D955.
29. Uhlén M, Fagerberg L, Hallström BM, et al. Tissue-based map of the human proteome. *Science.* 2015;347:1260419.
30. Veeramachaneni V, Makalowski W, Galdzicki M, Sood R, Makalowska I. Mammalian overlapping genes: the comparative perspective. *Genome Res.* 2004;14:280-286.
31. Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res.* 2019;47:D419-D426.
32. Schoch CL, Ciuffo S, Domrachev M, et al. NCBI taxonomy: a comprehensive update on curation, resources and tools. *Database.* 2020;2020:baaa062.
33. Released 2021. *IBM SPSS Statistics for Windows, Version 28.0.* IBM Corp; 2021.
34. Tóth G, Gáspári Z, Jurka J. Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res.* 2000;10:967-981.
35. Gardiner-Garden M, Frommer M. CpG islands in vertebrate genomes. *J Mol Biol.* 1987;196:261-282.
36. Akan P, Deloukas P. DNA sequence and structural properties as predictors of human and mouse promoters. *Gene.* 2008;410:165-176.
37. De Ferrari L, Aitken S. Mining housekeeping genes with a naive Bayes classifier. *BMC Genomics.* 2006;7:277.
38. Lovegrove BG. A phenology of the evolution of endothermy in birds and mammals. *Biol Rev Camb Philos Soc.* 2017;92:1213-1240.
39. Farmer CG. Parental care: the key to understanding endothermy and other convergent features in birds and mammals. *Am Nat.* 2000;155:326-334.
40. Harding L, Jackson A, Barnett A, et al. Endothermy makes fishes faster but does not expand their thermal niche. *Funct Ecol.* 2021;35:1951-1959.
41. Morris A. Obesity: olfactory senses linked to metabolism. *Nat Rev Endocrinol.* 2017;13:499.
42. Riera CE, Tsaousidou E, Halloran J, et al. The sense of smell impacts metabolic health and obesity. *Cell Metab.* 2017;26:198-211.e5.
43. Schöne S, Jurk M, Helabad MB, et al. Sequences flanking the core-binding site modulate glucocorticoid receptor structure and activity. *Nat Commun.* 2016;7:12621.
44. Pin-on P, Aporntewan C, Siriluksana J, Bhummaphan N, Chanvorachote P, Mutirangura A. Targeting high transcriptional control activity of long mononucleotide A-T repeats in cancer by Argonaute 1. *Gene.* 2019;699:54-61.
45. Bhummaphan N, Pin-on P, Phiboonchaiyanan PP, et al. Targeting multiple genes containing long mononucleotide A-T repeats in lung cancer stem cells. *J Transl Med.* 2021;19:231.