Enhanced Functional Potential of Pseudogene-associated lncRNA Genes in Mammals

Ze-Hao Zhang[1,#], Bo-Han Li[1,#], Yin-Wei Wang[1], Sheng Hu Qian[1], Lu Chen[1], Meng-Wei Shi[1], Hao Zuo[1,*], Zhen-Xia Chen[1,2,3,*]

[1]*Hubei Hongshan Laboratory, Hubei Key Laboratory of Metabolic Abnormalities and Vascular Aging, Hubei Key Laboratory of Agricultural Bioinformatics, College of Life Science and Technology, College of Biomedicine and Health, Interdisciplinary Sciences Institute, Huazhong Agricultural University, Wuhan 430070, PR China*

[2]*Shenzhen Institute of Nutrition and Health, Huazhong Agricultural University, Shenzhen 518000, PR China*

[3]*Shenzhen Branch, Guangdong Laboratory for Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518000, PR China*

[#] Equal contribution.

[*] Corresponding author.

E-mail: zhen-xia.chen@mail.hzau.edu.cn (Chen Z); zuohao@webmail.hzau.edu.cn (Zuo H).

Total word counts (from "Introduction" to "Materials and methods"): 8142

Total reference counts: 86

Total figures: 7

Total tables: 0

Total supplementary figures: 26

Total supplementary tables: 7

Total letter counts in the article title: 70

Total letter counts in the running title: 43

Total keyword counts: 4

Total word counts in Abstract: 191

Abstract

The functional significance of long non-coding RNAs (lncRNAs) remains a subject of debate, largely due to the complexity and cost associated with their validation experiments. However, emerging evidence suggests that pseudogenes, once viewed as genomic relics, may contribute to the origin of functional lncRNA genes. In this study spanning eight species, we systematically identified pseudogene-associated lncRNA genes using our PacBio long-read sequencing data and published RNA-seq data. Our investigation revealed that pseudogene-associated lncRNA genes exhibit heightened functional attributes compared to their non-pseudogene-associated counterparts. Notably, these pseudogene-associated lncRNAs show protein-binding proficiency, positioning them as potent regulators of gene expression. In particular, pseudogene-associated sense lncRNAs retain protein-binding capabilities inherited from parent genes of pseudogenes, thereby demonstrating greater protein-binding proficiency. Through detailed functional characterization, we elucidated the unique advantages and conserved roles of pseudogene-associated lncRNA genes, particularly in the context of gene expression regulation and DNA repair. Leveraging cross-species expression profiling, we demonstrated the prominent contribution of pseudogene-associated lncRNA genes to aging-related transcriptome changes across nine human tissues and eight mouse tissues. Overall, our findings demonstrate enhanced functional attributes of pseudogene-associated lncRNA genes and shed light on their conserved and close association with aging.

KEYWORDS: long non-coding RNA; pseudogene; functionality; aging

Introduction

Long non-coding RNAs (lncRNAs), defined as RNA molecules exceeding 200 nucleotides (nt) lacking protein-coding potential, have garnered significant interest with the advances in transcriptome sequencing technology. Although tens of thousands of lncRNA genes have been identified in mammalian genomes [1], only a small fraction have been functionally characterized [2,3]. The majority of lncRNAs are still considered as transcriptional by-products [4,5], and the proportion of functional lncRNA genes in the total lncRNA gene collection remains controversial [6]. Meanwhile, function assays for lncRNA genes are usually of low throughput, high complexity and high cost, hindering comprehensive exploration of their functional roles.

Pseudogenes, characterized as defective copies of protein-coding genes [7], are widely distributed in mammalian genomes [8]. Pseudogenes can be classified into multiple classes based on their mechanism of origin [9]. Duplicated pseudogenes arise from duplications at the DNA level, such as whole genome duplications and chromosomal segment duplications, whereas processed pseudogenes arise from the integration of cDNAs formed by reverse transcription of mRNAs into the genome [10]. These two classes of pseudogenes represent the two major categories of pseudogenes. Although pseudogenes exhibit sequence similarity to their corresponding parent genes, they are usually not transcribed or do not produce normal mRNA due to varying degrees of mutations in regulatory elements or coding regions [11]. Therefore, pseudogenes have long been considered as non-functional "genomic fossils" [12,13].

However, recent studies have unveiled the potential of pseudogenes as precursors for functional lncRNA genes [14]. For example, the mouse pseudogene *Rps15a-ps4* has been shown to transcribe the lncRNA Lethe [15], which regulates the transcription of downstream target genes by affecting the DNA-binding activity of the RelA subunit of the transcription factor NF-κB. The lncRNA produced by the *Oct4p4* pseudogene regulates the expression level of its parent gene *Oct4* through epigenetic modifications, thus playing an important role in stem cell differentiation [16]. In addition, lncRNAs transcribed from the *AGPG* pseudogene have been reported to promote glycolysis and

88 cell cycle progression in human [17]. The lncRNAs produced by the pseudogenes
89 *PCNAP1* and *PTENP1* play a role in the development of hepatocellular carcinoma and
90 aortic aneurysm by regulating the expression levels of the parent genes *PCNA* and
91 *PTEN*, respectively [18,19]. Together, pseudogenes can be transcribed into lncRNAs,
92 which are involved in various regulatory functions. Accordingly, pseudogenes, which
93 are often regarded as non-functional remnants of coding genes, may represent an
94 appropriate substrate for the origin of functional lncRNA genes.

95 Here we identified pseudogene-associated lncRNA genes in vertebrates using
96 PacBio full-length sequencing. Our findings indicate an enrichment of lncRNA genes
97 at pseudogenes, suggesting that pseudogenes may serve as favorable substrates for the
98 origin of functional lncRNA genes. Further characterization of pseudogene-associated
99 lncRNA genes revealed enhanced functional attributes and protein-binding proficiency.
100 These lncRNA genes exhibited unique advantages and conserved roles in gene
101 expression regulation and DNA repair. Importantly, we demonstrated the prominent
102 contribution of pseudogene-associated lncRNA genes to the aging process across nine
103 human tissues at the single-cell level and confirmed their conserved association with
104 aging in mouse. Taken together, our study sheds light on the heightened functional
105 attributes of pseudogene-associated lncRNA genes and their potential implications in
106 aging.

107

108 ## Results

109 ### Identification of pseudogene-associated lncRNA genes across eight vertebrate
110 genomes

111 Long-read transcriptome sequencing has proven to be accurate and effective in
112 improving genome annotation and discovering new genes [20,21]. To accurately
113 characterize lncRNA genes, we performed PacBio isoform sequencing (Iso-seq), as
114 well as RNA-seq to correct for Iso-seq data, for the adult female and male tissues (brain,
115 heart, gut, and gonad) of three representative vertebrate models, including zebrafish

116 (teleost), chicken (bird), and mouse (mammal) (**Figure 1**A). As expected, the
117 transcripts detected by PacBio Iso-seq were longer in length and had a higher number
118 of exons than the Ensembl annotated transcripts (Figure 1B, C), suggesting that our
119 home-made datasets produced high-quality genome annotations. To broaden the scope
120 for comparative studies, we also included five additional mammals (human, macaque,
121 rat, rabbit, and opossum), and a total of eight species were covered for subsequent
122 analysis. To ensure comprehensive lncRNA annotation, we integrated the Ensembl
123 lncRNA annotation with novel lncRNAs from our home-made Iso-seq datasets and
124 published lncRNA sets from previous studies (Methods; Figure S1), resulting in a
125 robust annotation of 9002 (in zebrafish) to 39,461 (in human) lncRNA genes for each
126 species (Figure 1D). Furthermore, novel lncRNA genes annotated by long-read
127 transcriptome sequencing frequently exhibit higher transcript length, higher sequence
128 conservation, and higher expression level in comparison with those annotated by short-
129 read sequencing (Figure S2A–C). These results suggest that long-read transcriptome
130 sequencing can facilitate the accurate identification of novel lncRNA genes with more
131 functional properties.

132 We then employed PseudoPipe, a computational pipeline based on homology search
133 [22], to identify pseudogene candidates across the eight genomes. Pseudogenes that
134 could not be reliably assessed as processed or duplicated pseudogenes were classified
135 as fragment pseudogenes. After rigorous filtering, 6058 (in chicken) to 19,843 (in
136 mouse) pseudogenes were retained in each species (Methods; Figure 1E). These
137 pseudogenes covered 84.2% (in human) and 81.6% (in mouse) of the Ensembl
138 pseudogene annotation, indicating a low false negative rate. As previous studies
139 [8,23,24], mammals have a higher abundance of pseudogenes, with processed
140 pseudogenes being more prevalent than duplicated pseudogenes. Conversely, in
141 zebrafish and chicken, the proportion of duplicated pseudogenes is higher than that of
142 processed pseudogenes (Figure S2D). Additionally, the number and density of
143 processed pseudogenes in chicken and zebrafish are lower than in mammals (Figure
144 S2E). Chicken CR1 reverse transcriptase is unlikely to copy polyadenylated mRNAs,

145 whereas polyadenylated tails are potential substrates for mammalian L1 reverse
146 transcriptase [25,26]. This may explain the differences in the abundance of processed
147 pseudogenes between chicken and mammals. In zebrafish, the genomic coverage of
148 retrotransposon elements is only a quarter of that in human, which could account for
149 the lack of processed pseudogenes [27].

150 Next, we identified the pseudogene-associated lncRNA genes based on two
151 different overlap patterns (Methods; Figure 1F, G; Table S1), and found 514 (in
152 chicken) to 2458 (in human) lncRNA genes, accounting for 2.4% to 8.3% of all the
153 lncRNA genes in each species. A robust positive correlation was identified between the
154 number of pseudogene-associated lncRNA genes and the total number of lncRNA genes
155 across diverse species (Figure S2F). This finding suggests that although part of the
156 variability in the number of pseudogene-associated lncRNA genes may be biologically
157 meaningful, much is likely to be explained by unequal size of the lncRNA gene
158 repertoire. Consequently, factors influencing the size of the lncRNA annotation,
159 including the quality of genome sequence and annotation, as well as the amount and
160 sequencing depth of RNA-seq data used to identify novel lncRNA genes [28], affect
161 the number of pseudogene-associated lncRNA genes. Additionally, 41.7% (in human)
162 to 75.2% (in rat) of the pseudogene-associated lncRNA genes in each species were
163 novel lncRNA genes (Figure S2G). This highlights the significance of comprehensive
164 lncRNA annotation. For example, zebrafish novel lncRNA gene *PB.7298* and mouse
165 novel lncRNA gene *PB.5227* detected by Iso-seq overlapped with the pseudogene and
166 were identified as pseudogene-associated lncRNA genes, with the first exon of both
167 derived from the pseudogene (Figure 1H). *PB.7298* was found to be expressed in the
168 gut and gonads, whereas *PB.5227* was only expressed in the testis.

169 **Enrichment of lncRNA genes at pseudogenes across species**

170 To understand the contribution of pseudogenes to the makeup and function of lncRNA
171 genes, we evaluated the enrichment of lncRNA genes at pseudogenes (Methods) and
172 found that both sense and antisense lncRNA genes were enriched at pseudogenes
173 compared to shuffled regions in mammals and zebrafish (empirical $P \leq 0.001$) (**Figure**

174 **2**A). After excluding the potential effects of chromosomal biases or repetitive
175 sequences (Methods), similar results were still observed (Figure 2A). We further
176 evaluated the enrichment of lncRNA genes at each class of pseudogenes, and found that
177 both duplicated and processed pseudogenes showed high enrichment of lncRNA genes
178 in comparison to shuffled regions (Figure S3). These findings imply that pseudogenes
179 act as relics of protein-coding gene copies to provide favorable substrates for the origin
180 of functional lncRNA genes, which are retained by purifying selection during evolution.

181 To further confirm whether the origin of lncRNA genes is functionally gained or
182 lost for pseudogene regions. We evaluated the sequence conservation of overlapping
183 and non-overlapping regions within the pseudogenes overlapping with sense and
184 antisense lncRNA genes using PhastCons score in human and mouse. We found that
185 regions within pseudogenes that overlap with exons of lncRNA genes are more
186 conserved than non-overlapping portions (two-sided Wilcoxon rank sum test, $*P < 0.05$,
187 $***P < 0.001$) (Figure 2B), suggesting that these pseudogene regions acquire enhanced
188 functional properties through the origin of sense or antisense lncRNA genes. For
189 example, within the pseudogene HOS_PG_6183, which overlaps with the human
190 lncRNA gene *ROCK1P1*, both PhastCons and PhyloP scores showed higher sequence
191 conservation in the regions overlapping the *ROCK1P1* exons (Figure 2C).

192 **Enhanced functional attributes of pseudogene-associated lncRNA genes in**
193 **sequence**

194 The enrichment of lncRNA genes at pseudogenes implies that pseudogenes may have
195 endowed pseudogene-associated lncRNA genes with functionality. To explore the
196 functional potential of pseudogene-associated lncRNA genes, we conducted
197 comparative analyses with non-pseudogene-associated (NPA) counterparts, focusing
198 on evolutionary age, sequence conservation, intrapopulation variation, and structural
199 attributes.

200 We estimated the evolutionary age of human and mouse lncRNA genes using 1:1
201 orthologous lncRNA genes based on the phylogenetic relationships of each species
202 (Methods; Table S2). Although large proportions of lncRNA genes were species-

specific in human (38.9%) and mouse (35.6%), both pseudogene-associated sense (PAS) and pseudogene-associated antisense (PAA) lncRNA genes had higher proportions of genes in older age groups, and lower proportions of genes in younger age groups than NPA lncRNA genes (**Figure 3**A; Figure S4A). Next, we analyzed the enrichment of age groups for each category of lncRNA genes using all lncRNA genes as background genes. Take human for example, we found that NPA lncRNA genes were not enriched in any of the evolutionary age groups (Figure 3B). In contrast, PAS lncRNA genes were enriched in 429 million years ago (Ma), and PAA lncRNA genes were enriched in 429 Ma, 319 Ma, and 160 Ma. A similar phenomenon was observed in mouse (Figure S4B). Taken together, our analysis suggests that pseudogene-associated lncRNA genes exhibit higher evolutionary conservation, which implies their greater functionality.

To further reveal the functional potential of pseudogene-associated lncRNA genes, we evaluated the cross-species DNA sequence conservation of these lncRNA genes (Figure 3C; Figure S4C; Table S2). We found that the exon sequence conservation of both PAS (median ~0.23 in human and ~0.24 in mouse) and PAA lncRNA genes (median ~0.18 in human and ~0.29 in mouse) was lower than that of the coding sequence (CDS) of protein-coding genes, but higher than that of NPA lncRNA genes (median ~0.07 in human and ~0.11 in mouse) and random intergenic regions (median ~0.07 in human and ~0.09 in mouse), and even higher than that of the 5' untranslated regions (UTRs) of protein-coding genes in both human and mouse (two-sided Wilcoxon rank sum test, $P < 0.001$). Conservation beyond the non-coding regions of the exons of protein-coding genes, demonstrating greater functionality of pseudogene-associated lncRNA genes. To ensure that the higher conservation of pseudogene-associated lncRNA genes compared to NPA counterparts was not due to their older age, we evaluated the conservation of lncRNA genes in each age group (Figure 3F; Figure S4D). We found that each category of lncRNA genes is highly conserved in the oldest age group, but both PAS and PAA lncRNA genes remain consistently more conserved than NPA counterparts in other age groups, suggesting stronger evolutionary constraints on pseudogene-associated lncRNA genes at each age.

Next, we investigated the intrapopulation variation of pseudogene-associated lncRNA genes in human [29]. Although the derived allele frequencies of single nucleotide polymorphisms (SNP) found in the exons of pseudogene-associated lncRNA genes were not different from those found in the exons of NPA lncRNA genes (two-sided unpaired T test, $P > 0.05$) (Figure S4E), the SNP density of exons of PAS lncRNA genes was lower than that of exons of NPA lncRNA genes as well as random intergenic regions (two-sided Wilcoxon rank sum test, $P < 0.001$) (Figure 3D). This suggests that PAS lncRNA genes may be subject to slightly stronger intrapopulation purifying selection than NPA lncRNA genes. In addition, the Combined Annotation-Dependent Depletion (CADD) score [30] is a deleteriousness score for human variants. We found that CADD scores for SNPs in exons of both PAS and PAA lncRNA genes were higher than those in exons of NPA lncRNA genes (two-sided Wilcoxon rank sum test, $P < 0.001$) (Figure S4F), suggesting that variants in pseudogene-associated lncRNA genes are more pathogenic than those in NPA lncRNA genes. These results again implied the potential functionality of pseudogene-associated lncRNA genes.

Since the functionality of genes is often related to the structural attributes of their transcripts [1,31], we examined the structural attributes of the transcripts of pseudogene-associated lncRNA genes (Table S2). We found that both PAS and PAA lncRNA genes had higher transcript lengths and exon numbers than NPA lncRNA genes (two-sided Wilcoxon rank sum test, $P < 0.001$; two-sided unpaired T test, $P < 0.001$) (Figure 3E; Figure S4G, H), suggesting that the transcripts of pseudogene-associated lncRNA genes may carry more RNA structural domains, thereby favoring their interactions with other nucleic acid or protein molecules [32].

Altogether, these findings in sequence revealed that pseudogene-associated lncRNA genes possess enhanced functional attributes compared to NPA counterparts.

Heightened functional attributes of pseudogene-associated lncRNA genes in expression

To further explore the functionality of pseudogene-associated lncRNA genes in expression, we conducted comparisons of their expression profiles and epigenetic

features, including chromatin accessibility, histone modifications, and DNA methylation, near transcription start sites (TSSs) with protein-coding genes and NPA lncRNA genes.

Utilizing RNA-seq data from 32 human tissues, we determined expression profiles across tissues and found that the expression level of both PAS and PAA lncRNA genes, compared with NPA lncRNA genes, was closer to protein-coding genes (two-sided Wilcoxon rank sum test, $P < 0.001$) (Figure 3G; Figure S5; Table S2). Moreover, the expression ratios of both PAS and PAA lncRNA genes were higher than that of NPA lncRNA genes, and closer to that of protein-coding genes at all expression cutoffs (Figure 3H) and in all 32 tissues (Figure 3I). Notably, PAS lncRNA genes show the highest expression activity. We also observed a conserved high expression activity of PAS lncRNA genes in mouse (Figure S6A-D). In addition, each category of lncRNA genes both showed tissue-specific expression patterns (Figure S6E; Table S2), supporting the high tissue-specificity of lncRNA gene expression [33,34]. Meanwhile, both PAS and PAA lncRNA genes (median: 7 and 6 tissues), as well as coding genes, exhibited broader tissue expression profiles than NPA lncRNA genes (median: 4 tissues) in human (two-sided Wilcoxon rank sum test, $P < 0.001$) (Figure S6E), demonstrating their potential functional versatility across tissues.

Chromatin accessibility plays a crucial role in regulating gene expression by influencing the accessibility of DNA to transcriptional machinery and regulatory proteins. Therefore, to elucidate the causes of active expression of pseudogene-associated lncRNA genes at the epigenomic level, we characterized chromatin accessibility near the transcription start sites (TSSs) in human. Compared to NPA lncRNA genes, both PAS and PAA lncRNA genes showed more housekeeping chromatin open regions by ATAC-seq, and PAS lncRNA genes also showed higher DNase accessibility by DNase-seq (**Figure 4**A, B), demonstating higher chromatin accessibility at proximal promoters of pseudogene-associated lncRNA genes.

Histone modifications may regulate gene expression by influencing chromatin structure and function. We found more signals near TSS of both PAS and PAA lncRNA

genes by ChIP-seq for histone modifications associated with open chromatin, including H3K4me3, H3K4me2, and H3K9ac, than that near TSS of NPA lncRNA genes (Figure 4C). PAS lncRNA genes also exhibited slightly higher H3K27ac signals compared to NPA lncRNA genes. These results are consistent with the higher chromatin accessibility near pseudogene-associated lncRNA genes. Moreover, we counted around the TSS the binding sites of ten chromatin modifying factors, including six components of the H3K4 methyltransferase Complex of Proteins Associated with Set1 (COMPASS) [35], two components of the histone acetyltransferase CBP/p300 complex [36], and two componets of the histone acetyltransferase GCN5/PCAF complex [37], and found that the binding levels of all the chromatin modifying factors except p300 were elevated in the proximal regions of both PAS and PAA lncRNA gene TSSs (Figure 4D), indicating the shaping of epigenomic signatures characteristic of enhanced transcriptional regulation. The higher signials of histone modifications associated with open chromatin and the higher binding levels to related chromatin modifiers near the TSS of pseudogene-associated lncRNA genes would allow their transcriptional activation.

Besides histone modifications, DNA methylation are also important regulators of gene expression. We thus examined DNA methylation in each 200 bp window within a total of 4kb regions upstream and downstream of the gene TSSs, and found that the DNA methylation of both PAS and PAA lncRNA genes was lower than that of NPA lncRNA genes in each window (two-sided Wilcoxon rank sum test, $P < 0.001$) (Figure 4E), and reached the lowest level at TSS, indicating that the pseudogene-associated lncRNA genes could more effectively avoid the silencing of gene expression caused by DNA methylation in the promoter region.

As an integration of chromatin accessibility, histone modifications, DNA methylation, and other epigenetic feature, chromHMM chromatin states serve as a molecular framework influencing gene expression patterns [38]. We found that the ratio of active promoter state regions in the proximal promoters of both PAS and PAA lncRNA genes was higher than that in the proximal promoters of NPA lncRNA genes as well as in the random intergenic region (two-sided Wilcoxon rank sum test, $P < 0.001$)

319 (Figure 4F). These results are consistent with their epigenomic characterization.

320 In summary, the characterization of gene expression, chromatin accessibility,
321 histone modifications, DNA methylation levels, and chromatin state collectively
322 indicated that pseudogene-associated lncRNA genes are subject to heightened levels of
323 active transcriptional regulation, thereby facilitating enhanced functionality.

324 **Protein-binding proficiency of pseudogene-associated lncRNAs hints they are**
325 **important regulators of gene expression**

326 lncRNAs usually function as regulators of gene expression [39]. In general, lncRNAs
327 often exert their regulatory roles by interacting with RNA-binding proteins (RBPs)
328 [40,41]. They regulate gene expression at the transcriptional level by recruiting
329 transcription factors and chromatin modifiers, or at the post-transcriptional level by
330 competing with mRNAs for binding to RNA stabilizing/destabilizing factors [2,42,43].
331 Therefore, to explore the potential function of pseudogene-associated lncRNA genes,
332 we scrutinized their features associated with RNA-protein binding in human.

333 N6-methyladenosine (m6A) is a prevalent and significant RNA modification known
334 to influence the binding between RNAs and proteins [44,45]. We investigated the
335 m6A modification in pseudogene-associated lncRNAs and found that the density of
336 m6A sites in both PAS and PAA lncRNAs was higher than that in NPA lncRNAs (two-
337 sided Wilcoxon rank sum test, $P < 0.001$) (**Figure 5**A; Table S2), suggesting a more
338 complex binding pattern between pseudogene-associated lncRNAs and proteins. Next,
339 we directly characterized the binding of pseudogene-associated lncRNAs to RBPs
340 (Table S2). In transcripts with RBP binding sites, the richness of RBPs bound to both
341 PAS and PAA lncRNAs was higher than those bound to NPA lncRNAs (two-sided
342 Wilcoxon rank sum test, $^{**}P < 0.01$, $^{***}P < 0.001$) (Figure 5B). Furthermore, the
343 diversity of RBPs bound to PAS lncRNAs was also higher than those bound to NPA
344 lncRNAs (Figure S7A). To exclude the effect of gene expression levels on the detection
345 of m6A sites as well as RBP binding sites, we evaluated these features associated with
346 RNA-protein binding in four expression level groups and found similar results (Figure
347 S7B-D). These results suggest that pseudogene-associated lncRNAs bind more

348 abundant RBPs than NPA lncRNAs, indicating that pseudogene-associated lncRNAs
349 are greater regulators of gene expression compared to NPA counterparts.

350 Although both PAS and PAA lncRNAs showed advantages in protein-binding
351 capacity, PAS lncRNAs exhibited higher m6A site density as well as more abundant
352 and diverse RBP binding sites compared to PAA lncRNAs (Figure 5A, B; Figure S7A).
353 Since PAS lncRNAs have homologous sequences of transcripts from parent genes of
354 pseudogenes, whereas PAA lncRNAs have potential complementary sequences of
355 transcripts from parent genes of pseudogenes, we hypothesized that the abundant RBP
356 binding sites of PAS lncRNAs are conferred by parent genes of pseudogenes. To test
357 this hypothesis, we calculated the Bray-Curtis similarity of RBP binding sites between
358 each category of pseudogene-associated lncRNAs and transcripts from parent genes of
359 pseudogenes. A higher RBP similarity was identified between PAS lncRNAs and
360 transcripts from parent genes of pseudogenes compared to gene pairs used as controls
361 (two-sided Wilcoxon rank sum test, $P < 0.001$) (Figure 5C). However, the RBP
362 similarity between PAA lncRNAs and transcripts from parent genes of pseudogenes did
363 not differ from that observed in gene pairs used as controls (Figure 5C). These findings
364 suggest that only PAS lncRNAs inherit the RBP binding sites from parent genes of
365 pseudogenes, thereby demonstrating greater proficiency in protein-binding.

366 Furthermore, we analyzed the enrichment of RBPs for each category of lncRNAs
367 using all lncRNAs as background. We found that NPA lncRNAs were not enriched for
368 RBPs, whereas PAS lncRNAs and PAA lncRNAs were enriched for 104 and 12 RBPs,
369 respectively (Figure S7E). These RBPs include a number of gene expression regulators
370 (Figure 5D). The top 10 RBPs enriched by PAS lncRNAs include transcriptional
371 regulators (ZNF662, GRWD1, BCLAF1, CHTOP, DDX3X), RNA splicing regulators
372 (PPIG, PRPF8, TRA2A), deubiquitinase (UCHL5), and mRNA stabilizing factor
373 (DDX24). Meanwhile, PAA lncRNAs were also enriched for DDX3X and UCHL5, as
374 well as other gene expression regulators such as regulators of RNA splicing and
375 processing (RBM15B, CSTF2, HNRNPL), translation regulator (LIN28B), and mRNA
376 degradation factors (KHSRP, YTHDF2). These results suggest that pseudogene-

377 associated lncRNAs may be involved in the regulation of gene expression at multiple
378 levels by binding to RBPs. Moreover, DDX24 [46] and DDX3X [47,48] were found to
379 be closely associated with DNA repair, indicating a potential involvement of
380 pseudogene-associated lncRNAs in DNA repair.

381 Co-expression can indicate functional relevance or regulatory relationships between
382 lncRNAs and protein-coding genes [28]. To infer the putative functions of pseudogene-
383 associated lncRNA genes, we identified co-expressed protein-coding genes for each
384 lncRNA gene in human and mouse (Methods). We further analyzed the enrichment of
385 co-expressed genes to identify co-expressed genes more closely associated with the
386 function of each category of lncRNA genes. We found that both PAS and PAA lncRNA
387 genes were enriched for more co-expressed protein-coding genes compared to NPA
388 lncRNA genes in both human and mouse (Figure S7F, G). Next, co-expressed genes
389 enriched by each category of lncRNA genes were analyzed for Gene Ontology (GO)
390 enrichment analysis. The co-expressed genes enriched by NPA lncRNA genes were not
391 enriched in any of the entries in human and were enriched in mitosis-related entries in
392 mouse (Figure S7H). Whereas the co-expressed genes enriched by PAS lncRNA genes
393 were enriched in functions related to the regulation of gene expression, such as
394 histone modification, chromatin organization, chromatin remodeling, mRNA metabolic
395 process, RNA catabolic process, and regulation of RNA stability in both human and
396 mouse (Figure 5E; Figure S7I). Meanwhile, the co-expressed genes enriched by PAA
397 lncRNA genes were enriched in similar functions related to gene expression regulation
398 in mouse (Figure S7J, K). These results again implies the important role of pseudogene-
399 associated lncRNA genes in gene expression regulation. Enrichment analysis also
400 revealed that pseudogene-associated lncRNA genes have conserved functions in DNA
401 repair, double-strand break repair, and regulation of response to DNA damage stimulus
402 (Figure 5E; Figure S7I, K).

403 Taken together, our analyses about RBPs and co-expression collectively suggested
404 that pseudogene-associated lncRNA genes have unique advantages and conserved
405 functions in gene expression regulation and DNA repair, which may explain the high

406 conservation of pseudogene-associated lncRNA genes.

407 **Deciphering aging-related expression changes of pseudogene-associated lncRNA**
408 **genes across tissues**

409 Aging is often accompanied by reprogramming of the transcriptome and
410 downregulation of DNA repair levels. Motivated by the putative functions of
411 pseudogene-associated lncRNA genes, we extended our scope to aging. To
412 comprehensively explore the aging-related expression alterations of pseudogene-
413 associated lncRNA genes, we used single-cell/nucleus RNA sequencing (sc/snRNA-
414 seq) data from 9 human tissues to characterize the expression profile of pseudogene-
415 associated lncRNA genes in young and old groups (Methods; Table S3).

416 After cell filtration, we retained a total of 850,860 high-quality single cells/nuclei
417 and identified 72 cell types based on canonical cell markers in all tissues (Figure S8;
418 Table S4). We found that both PAS and PAA lncRNA genes were more actively
419 expressed than NPA lncRNA genes in each cell type of each tissue (two-sided Fisher's
420 exact test, $P < 0.001$) (Figure S9, 10). Meanwhile, both PAS and PAA lncRNA genes,
421 along with coding genes, displayed more extensive cell type expression profiles than
422 NPA lncRNA genes across all tissues (two-sided Wilcoxon rank sum test, $P < 0.001$)
423 (Figure S11), indicating their potential functional versatility across cell types.

424 Subsequently, a differential expression analysis was conducted between the young
425 and old age groups in each cell type of each tissue (Table S5). We found that the
426 proportion of aging-related differentially expressed genes (DEGs) in both PAS and PAA
427 lncRNA genes was higher than that in NPA lncRNA genes in each tissue (**Figure 6**A,
428 B), suggesting a more prominent contribution of pseudogene-associated lncRNA genes
429 to aging-related transcriptome changes and their closer association with aging.
430 Furthermore, we also observed that the close association of pseudogene-associated
431 lncRNA genes with aging was cell type-specific (Figure S12). For instance, in the brain,
432 the proportion of up-regulated genes in PAS lncRNA genes was higher than that in NPA
433 lncRNA genes in 6 out of 8 cell types, suggesting that pseudogene-associated lncRNA
434 genes respond differently to aging in different cell types of the same tissue. This

underscores the necessity for a detailed examination of their aging-related alterations in expression at the single-cell level.

Among aging-related DEGs, we observed that the proportion of tissue-shared DEGs (genes differentially expressed in at least two tissues) in both PAS and PAA lncRNA genes was higher than that in NPA lncRNA genes (two-sided Fisher's exact test, $P < 0.001$) (Figure 6C), suggesting that pseudogene-associated lncRNA genes respond to aging more broadly across tissues. Considering the broader tissue expression profile of pseudogene-associated lncRNA genes (Figure S6E, S13), we proceeded to investigate the number of tissues in which the genes exhibited aging-related differential expression in each expression breadth group. Both PAS and PAA lncRNA genes were found to exhibit aging-related differential expression in a greater number of tissues, particularly in the group of genes expressed in nine tissues (Figure 6D), suggesting that pseudogene-associated lncRNA genes are more commonly involved in the aging process. One illustrative example of pseudogene-associated lncRNA genes that exhibit aging-related differential expression in multiple tissues is *JPX*, which demonstrates aging-related differential expression in 20 cell types across 8 tissues (Figure 6E). *JPX* has been demonstrated to be directly related to aging and represents a potential therapeutic target for the treatment of age-related atherosclerosis [49]. Overall, these findings indicate that pseudogene-associated lncRNA genes exhibit a positive and extensive response to aging.

To investigate whether pseudogene-associated lncRNA genes show a close association with aging due to transcriptional regulation, we assessed the enrichment of transcription factor (TF) binding sites in the proximal promoters of each category of lncRNA genes in human. Both PAS and PAA lncRNA genes were enriched for the binding sites of aging-related TFs (Figure S14), including NRF1 [50], E2F6, RB1 [51], PML [52], and SREBF2 [53]. Furthermore, the binding sites of longevity factors (FOXO3, SIRT6) [54,55], rejuvenation factor (SOX5) [56], and aging modulators (MTOR, ZBTB7A, CXXC1, SREBF1) [57,58] were also enriched in the proximal promoters of PAS lncRNA genes. Conversely, NPA lncRNA genes were not enriched

464 for the binding sites of these aging-related TFs. These results suggest that pseudogene-
465 associated lncRNA genes are subject to transcriptional regulation of aging-related TFs,
466 thereby closely associating with aging.

467 To further explore the putative functions of pseudogene-associated lncRNA genes
468 in aging, we conducted GO enrichment analysis based on the guilt-by-association
469 method for aging-related DEGs in each category of lncRNA genes. The aging-related
470 DEGs in both PAS and PAA lncRNA genes were enriched for aging-related biological
471 processes (Figure 6F), including regulation of inflammatory response, NF-kappaB
472 signaling, TOR signaling, response to oxidative stress, cellular senescence, DNA repair,
473 and telomere maintenance. In contrast, the aging-related DEGs in NPA lncRNA genes
474 were not enriched for these aging-related entries. These findings suggest that
475 pseudogene-associated lncRNA genes regulate aging-related pathways in aging.

476 To investigate whether the close association of pseudogene-associated lncRNA
477 genes with aging is conserved in mouse, we used RNA-seq data from 8 mouse tissues
478 to characterize the aging-related expression changes of pseudogene-associated lncRNA
479 genes (Methods; Table S6). We found that both PAS and PAA lncRNA genes showed a
480 higher proportion of aging-related DEGs and a broader aging-related differential
481 expression compared to NPA lncRNA genes across mouse tissues (Figure S15). To
482 experimentally validate the aging-related expression changes of pseudogene-associated
483 lncRNA genes, we quantified the expression levels of the top 3 PAS and PAA lncRNA
484 genes differentially expressed in the most mouse tissues by quantitative real-time PCR
485 (RT-qPCR). These pseudogene-associated lncRNA genes showed aging-related
486 differential expression in multiple tissues by both RNA-seq and RT-qPCR analysis
487 (**Figure 7**A, B, C; Figure S16). These results confirm the conserved and close
488 association between pseudogene-associated lncRNA genes and aging. Furthermore, we
489 identified two orthologous PAS lncRNA gene pairs and two orthologous PAA lncRNA
490 gene pairs that exhibited aging-related differential expression in both human and mouse
491 (Figure 7D; Figure S17). These lncRNA genes may play a conserved role in aging.
492 Among them, *PVT1* has been shown to be involved in the regulation of aging-related

493 diseases and skin aging [59,60]. We validated aging-related differential expression of
494 mouse orthologs in these orthologous gene pairs by RT-qPCR (Figure 7E). Together,
495 the characterization of aging-related expression changes across tissues and species
496 suggested that the association between pseudogene-associated lncRNA genes and aging
497 is tight and conserved. Therefore, further rigorous studies of the pseudogene-associated
498 lncRNA genes identified in this study will facilitate the discovery of novel aging
499 markers and therapeutic targets for aging-related diseases in the future.
500

501 **Discussion**

502 lncRNA-related research is growing by the day, but there is still an intense debate about
503 the proportion of functional lncRNA genes in the total lncRNA gene collection [6].
504 Studies of eukaryotic evolutionary dynamics and transcription machinery have shown
505 that lncRNA annotation contain a large number of non-functional lncRNA genes [61].
506 The co-occurrence of non-functional junk and functional lncRNA genes in lncRNA
507 annotation hinders the broader exploration of functional lncRNA genes. Pseudogenes
508 are often considered as non-functional sequences and thus excluded from functional
509 genomics studies. However, increasing evidence suggests that pseudogenes may be
510 favorable substrates for the origin of functional lncRNA genes [16,19,62].

511 We identified pseudogene-associated lncRNA genes using genome-wide
512 pseudogenes and a comprehensive lncRNA annotation based on the Ensembl lncRNA
513 annotation, as well as novel lncRNA genes from our home-made Iso-seq datasets and
514 published lncRNA sets from previous studies. More than 40% of pseudogene-
515 associated lncRNA genes are novel lncRNA genes in each species, indicating the
516 significance of a comprehensive lncRNA gene repertoire. Our human lncRNA gene
517 annotation is not as extensive as that provided by the LncBook database, which offers
518 a comprehensive and high-quality collection of human lncRNA genes [63] (Figure S18).
519 However, this limitation arises from the inclusion of eight species in our study. While
520 ensuring systematic construction of the lncRNA gene annotation and comparability
521 across different species, the comprehensiveness of our lncRNA annotation is

522 comparable to that found in recent studies [31,41]. Future enhancements in lncRNA
523 gene annotation across various species will facilitate the identification of pseudogene-
524 associated lncRNA genes.

525 Our analysis suggests that pseudogenes are favorable substrates for origin of
526 functional lncRNA genes in mammals. This notion encompasses two potential
527 mechanisms: (1) the formation of pseudogenes in intergenic regions, resulting in the
528 origin of novel lncRNA genes, and (2) the integration of pseudogenes into existing
529 genes, leading to the origin of novel chimeric genes that transcribe lncRNAs.
530 Collectively, our findings provide new insights into the origin of lncRNA genes and
531 support the view that new genes can arise from DNA-based or RNA-based gene
532 duplication [64].

533 Multidimensional evidence supports the enhanced functionality of pseudogene-
534 associated lncRNA genes, including higher evolutionary conservation, higher exon
535 sequence conservation, longer transcripts, more active expression, more active
536 transcriptional regulation, and greater protein-binding capabilities compared to NPA
537 counterparts. In order to provide a more accurate reflection of the influence of
538 pseudogenes on lncRNA genes, a filtering process was conducted on pseudogene-
539 associated lncRNA genes based on stricter overlap length thresholds between
540 pseudogenes and lncRNA genes. We found that pseudogene-associated lncRNA genes
541 showed further heightened functional potential at stricter overlap length cutoffs (Figure
542 S19–22). Taken human PAS lncRNA genes for example, they exhibit invariant
543 evolutionary age distribution and RBP-binding capabilities as well as higher
544 conservation, transcript length, expression levels, and m6A site density at stricter
545 overlap length thresholds (Figure S19). Additionally, transcriptional activity might
546 enhance the identification criteria of pseudogene-associated lncRNA genes. However,
547 the functional potential of pseudogene-associated lncRNA genes is invariant at stricter
548 expression level cutoffs (Figure S23-26). Altogether, these results ensure the reliability
549 of enhanced functional attributes of pseudogene-associated lncRNA genes.

550 Both PAS and PAA lncRNA genes show an advantage in serving as regulators of

551 gene expression over NPA lncRNA genes. However, PAS and PAA lncRNA genes
552 exhibit distinct functional advantages. Previous studies [62] have implied that PAS
553 lncRNA genes may inherit the RBP binding sites from parent genes of pseudogenes,
554 and our analysis confirmed this view. Therefore, PAS lncRNA genes have greater
555 protein-binding proficiency compared to PAA lncRNA genes. Conversely, PAA
556 lncRNA genes have an advantage in binding to the mRNAs of parent genes of
557 pseudogenes [65,66].

558 Furthermore, a transcriptome profiling across nine human tissues from young and
559 old age groups at the single-cell level revealed that pseudogene-associated lncRNA
560 genes exhibited prominent and widespread aging-related expression alterations. These
561 pseudogene-associated lncRNA genes are regulated by aging-related TFs and regulate
562 aging-related pathways. Their close association with aging was further confirmed to be
563 conserved in mouse by RNA-seq data and RT-qPCR. Taken together, pseudogene-
564 associated lncRNA genes will constitute a significant resource for the mining of
565 functional lncRNA genes and provide informative insights into human aging biology in
566 the future.

567

568 Materials and methods

569 Identification of Pseudogenes

570 We downloaded genomic data from Ensemble (release-108) for human GRCh38,
571 macaque Mmul_10, mouse GRCm39, rat mRatBN7.2, rabbit OryCun2.0, opossum
572 ASM229v1, chicken GRCg7b, and zebrafish GRCz11, including genome sequences
573 with the repetitive sequences masked, known protein sequences, and genome
574 annotations. The software PseudoPipe used these genomic data to identify genome-
575 wide pseudogene candidates as well as their classification and corresponding parent
576 genes [22]. To eliminate false positives, we performed quality control on pseudogene
577 candidates (identity ≥ 20%; match length ≥ 10% of the query sequence; match length ≥
578 30 amino acids; E-value ≤ 1E-5). High-quality pseudogenes that passed quality control
579 were used for downstream analysis.

580 **Construction of lncRNA Annotation**

581 We complemented the Ensemble lncRNA annotation with long-read transcriptome
582 sequencing data. We first processed mouse, chicken, and zebrafish Iso-seq data
583 generated in our lab. The mouse data have been published [14] and deposited in the
584 GEO database as GSE176018. GSE176018 contains RNA-seq and Iso-seq data for six
585 tissues (including brain, cerebellum, heart, colon, ovary, and testis) from adult
586 C57BL/6J female and male mice. Chicken and zebrafish data include RNA-seq and Iso-
587 seq data for six tissues (including brain, cerebellum, heart, colon, ovary, and testis) from
588 adult female and male chickens and five tissues (including brain, heart, colon, ovary,
589 and testis) from adult female and male zebrafish. Raw Iso-seq data were processed by
590 CCS (v6.2.0) to generate circular consensus sequences (parameters: -j 64 --min-passes
591 1 --min-length 300 --min-rq 0.9). Circular consensus sequences were processed by lima
592 (v2.3.99) to remove primer sequences to produce full-length sequences (parameters: -j
593 64 --isoseq). Full-length non-chimeric sequences were generated by Isoseq3 refine
594 (v3.5.0) (parameters: -j 64 --require-polya). Full-length transcript sequences were
595 generated by Isoseq3 cluster (v3.5.0). RNA-seq data were used by LoRDEC (v0.9) to
596 correct full-length transcript sequences [67]. Corrected transcript sequences longer than
597 200 were compared to the genome using Minimap2 (v2.26-r1175) (parameters: -t 64 -
598 ax splice --secondary=no -C5 -O6,24 -B4 -uf) [68]. Redundant isoforms were collapsed
599 using the script collapse_isoforms_by_sam.py from cDNA_Cupcake (v29.0.0;
600 https://github.com/Magdoll/cDNA_Cupcake) (parameters: --dun-merge-5- shorter -c
601 0.85 -i 0.95 --cpus 28). 5' degraded transcripts were removed using
602 get_abundance_post_collapse.py and filter_away_subset.py. The above process
603 completed the transcriptome annotation based on long-read transcriptome sequencing
604 data in mouse, chicken, and zebrafish. We obtained human transcriptome annotations
605 based on long-read transcriptome sequencing data from GTEx V9 release.

606 Bedtools (v2.30.0) was used to select novel genes that did not overlap with
607 Ensemble genes other than pseudogenes from the above annotations based on long-read
608 transcriptome sequencing. We then removed all genes with evidence of coding potential.

609 CPC2 (v0.1) [69], CNCI (v2) [70], PLEK (v1.2) [71], and LGC (v1.0) [72] were used
610 to detect protein-coding potential. InterProScan (v5.61-93.0) [73] was used to detect
611 transcripts with the Pfam protein domains. Only genes that successfully passed all five
612 filters for all their transcripts were identified as novel lncRNA genes based on long-
613 read transcriptome sequencing.

614 Next, we obtained high-quality lncRNA gene datasets annotated with short-read
615 transcriptome sequencing data based on multiple developmental stages in seven tissues
616 for seven species (human, rhesus macaque, mouse, rat, rabbit, opossum, and chicken)
617 from the previous study by Kaessmann et al [1]. We also obtained a high-quality
618 lncRNA gene dataset from the FishGET database for zebrafish, annotated based on
619 short-read transcriptome sequencing data [74]. These lncRNA gene dataset were
620 mapped to the genome used in this study using liftOver, and transcripts with changes
621 in transcript structure after mapping were excluded. lncRNAs that overlapped with
622 Ensemble genes other than pseudogenes and novel lncRNAs annotated based on long-
623 read transcriptome sequencing were also eliminated. The remaining lncRNAs were then
624 used as novel lncRNAs annotated based on short-read transcriptome sequencing. These
625 novel lncRNA annotations were merged with the Ensembl annotations as complements.

626 **Identification of Pseudogene-associated lncRNA genes**

627 We identified two categories of pseudogene-associated lncRNA genes based on two
628 different overlap patterns. Among the total lncRNA genes, we considered genes whose
629 exons overlap with pseudogenes on the same strand as PAS lncRNA genes, and genes
630 whose exons overlap with pseudogenes on the opposite strand as PAA lncRNA genes.
631 All remaining lncRNA genes were classified as NPA lncRNA genes. The data for the
632 two categories of pseudogene-associated lncRNA genes are provided in Table S1 and
633 Table S2.

634 **Estimation of Enrichment of lncRNA Genes at Pseudogenes**

635 To evaluate the enrichment of lncRNA genes at pseudogenes, Bedtools shuffle (v2.30.0)
636 was used to randomly generate 1000 sets of shuffled regions with the same number and
637 equal lengths as pseudogenes. Shuffled regions were not to be placed in exons of

638 protein-coding genes. Bedtools intersect was used to screen for overlap between
639 pseudogenes or shuffled regions and exons of lncRNA genes. Fold enrichment is a ratio
640 of an observed overlap compared to the expected overlap based on shuffling. The
641 number of pseudogenes overlapping exons of lncRNA genes was calculated as the
642 observed overlap. The mean of the number of shuffled regions overlapping exons of
643 lncRNA genes in each set of shuffled regions was calculated as the expected overlap.
644 The empirical p-value of enrichment is defined as the number of shuffled region sets
645 that show an equal or greater overlap than the observed overlap. To exclude the
646 potential effects of chromosomal biases, we utilized the Bedtools shuffle parameter "-
647 chrom" to make the shuffled regions exhibit the same chromosomal distribution as the
648 pseudogenes. To exclude the effects of repetitive sequences, we downloaded
649 annotations of repetitive sequences from UCSC Genome Browser and excluded
650 pseudogenes and lncRNA genes overlapping with repetitive sequences in the
651 enrichment analysis. We utilized the Bedtools shuffle parameter "-excl" to ensure that
652 the shuffled regions are not placed in repetitive sequences.

653 **Identification of Homologous lncRNA Genes and Estimation of lncRNA Gene Age**

654 For human and mouse lncRNA genes, we first used liftOver to project their genomic
655 coordinates to other species based on chain files generated by the UCSC Genome
656 Browser (parameter: -minMatch=0.1). We then applied Bedtools intersect to obtain the
657 intersection of lncRNA genes around 10 kb and identify syntenic lncRNA genes across
658 species. For each pair of syntenic lncRNA genes, we used gffread (v0.12.7) to obtain
659 the transcript sequences. Lastz (v1.04.22) was used to obtain the maximum alignment
660 score. We identified the one-to-one orthologous lncRNA genes based on the alignment
661 scores and the reciprocal-best rule. Finally, we used one-to-one orthologous lncRNA
662 genes to infer the evolutionary age of each lncRNA gene based on the phylogenetic
663 relationships of the species in which each lncRNA gene was transcribed, using the
664 parsimony rule. The ages of lncRNA genes with cross-species syntenic lncRNA genes,
665 but no one-to-one orthologous lncRNA genes, were categorized as "ambiguous".

666 **Sequence Conservation and Variation Analysis**

667 Genome-wide nucleotide resolution PhastCons scores were downloaded from the
668 UCSC Genome Browser. Average PhastCons scores for exonic regions were calculated
669 using bigWigAverageOverBed. Human common (minor allele frequency (MAF) ≥ 0.01)
670 single nucleotide polymorphism (SNP) data were downloaded from the dsSNP database.
671 The number of SNPs in exon regions was obtained by Bedtools intersect. We obtained
672 the derived allele frequencies of SNPs from phase 3 of the 1000 Genomes Project. To
673 avoid the bias caused by the high mutation rate in the CpG island region on the analysis,
674 we removed the CpG island region in the analysis of SNP derived allele frequency and
675 density. The CADD scores of the SNPs were obtained from the CADD database
676 (https://cadd.gs.washington.edu/).

677 **Expression Analysis**

678 We downloaded RNA-seq data from E-MTAB-2836 for a total of 200 samples from 32
679 different human tissues. C57BL/6J male and female adult mouse multiple tissues
680 (including brain, cerebellum, heart, colon, and gonad) RNA-seq data were obtained
681 from a previous study in our laboratory and stored in GSE176018. Raw FASTQ files
682 were quality controlled using FastQC (v0.11.9) and Trim Galore (v0.6.7), and reads
683 were aligned to the reference genome using STAR (v2.7.10b) [75] (parameters: --
684 runMode alignReads --runThreadN 28 --outSAMtype BAM SortedByCoordinate --
685 readFilesCommand zcat --sjdbOverhang 149 --twopassMode Basic --
686 outBAMsortingThreadN 20 --outSAMmultNmax 1 --outSJfilterReads Unique --
687 outSAMattrIHstart 0 --outSAMstrandField intronMotif). Reads with mapping scores
688 (MAPQ) < 20 were filtered out using SAMtools view (v1.16.1) [76]. Finally, gene
689 expression was quantified using featureCounts (v2.0.3) [77] and gene expression was
690 quantified using transcripts per kilobase million (TPM) to normalize gene expression
691 levels. Genes expressed in any sample (TPM ≥ 1) were used for subsequent analysis.

692 **Characterization of The Epigenome of Pseudogene-associated lncRNA Genes**

693 We obtained human epigenomic data from the NIH Roadmap Epigenomics Mapping

694 Consortium (http://egg2.wustl.edu/roadmap/). We downloaded DHS data for 53
695 consolidated epigenomes, H3K27ac peak data for 98 consolidated epigenomes,
696 H3K9ac peak data for 62 consolidated epigenomes, H3K4me2 peak data for 24
697 consolidated epigenomes, H3K4me3 peak data for 127 consolidated epigenomes,
698 fractional methylation data from WGBS for 37 consolidated epigenomes, and
699 chromHMM chromatin status for 127 consolidated epigenomes. In addition, we
700 obtained human housekeeping consensus peaks from a study based on the analysis of
701 1785 ATAC-seq and 231 scATAC-seq datasets [78]. We calculated the distribution of
702 housekeeping consensus peaks from ATAC-seq, DHS (DNase I hypersensitivity sites),
703 and four histone modification peaks in a total of 4kb regions upstream and downstream
704 of the gene TSSs. The distribution of each peak relative to the transcription start site
705 (TSS) was generated by deepTools [79], and the average fractional methylation of the
706 region around the TSS was calculated by bigWigAverageOverBed. Genome-wide
707 binding sites for TFs based on ChIP-seq data were downloaded from the GTRD
708 database [80]. We identified the proximal promoter of each gene (±500 bp from the
709 TSS) according to the criteria of the FANTOM5 consortium. Among the proximal
710 promoters with active promoter states, we counted the ratio of active promoter
711 chromHMM state regions.

712 **Characterization of The Features Associated with RNA-protein Binding**

713 N6-methyladenosine (m6A) site data based on m6A-seq and MeRIP-seq data were
714 downloaded from the REPIC database [81]. Transcriptome-wide binding site data for a
715 large number of RNA-binding proteins based on CLIP-seq were downloaded from the
716 POSTAR3 database [82]. We determined the binding ratio (binding length/total exon
717 length) of each RBP on each gene, and RBP richness was defined as the number of
718 types of different RBPs. The Shannon diversity index and Bray-Curtis similarity of
719 RBPs were all calculated using the R package vegan (v2.6-2). In order to evaluate RBP
720 similarity between each category of pseudogene-associated lncRNAs and transcripts
721 from parent genes of pseudogenes, two types of gene pairs were utilized as controls:
722 pseudogene-associated lncRNA genes and random protein-coding genes exhibiting a

723 matched expression level with parent genes of pseudogenes, as well as parent genes of
724 pseudogenes and random lncRNA genes exhibiting a matched expression level with
725 pseudogene-associated lncRNA genes.

726 **Screening of Co-expressed Genes for lncRNA Genes**

727 We calculated Spearman expression correlations and adjusted p-values (Holm's
728 multiple comparison test) between each lncRNA gene and each protein-coding gene.
729 To reduce false positives, we downloaded RNA-seq data from E-MTAB-6814 for a total
730 of 313 samples from 7 tissues and 23 developmental stages in human and from E-
731 MTAB-6798 for a total of 317 samples from 7 tissues and 14 developmental stages in
732 mouse. These data, along with those previously used for expression analysis, were used
733 for screening of co-expressed genes for lncRNA genes. Given the accessible chromatin
734 context and the disproportionate RNA expression observed in adult testis, testis samples
735 were excluded from this analysis. We identified co-expressed gene pairs as those with
736 correlation coefficients greater than 0.75 and adjusted p-values less than 0.001. We
737 further analyzed the enrichment of co-expressed genes for each category of lncRNA
738 genes using all lncRNA genes as background. The co-expressed genes significantly
739 enriched (two-sided Fisher's exact test, $P < 0.05$) by each category of lncRNA genes
740 were analyzed for the enrichment of GO entries. GO enrichment analysis using protein-
741 coding genes co-expressed with at least one lncRNA gene as background. Enrichment
742 analysis was performed using clusterProfiler (v4.8.1) [83].

743 **Pre-processing and Analysis of sc/snRNA-seq Data**

744 We downloaded sc/snRNA-seq data from GSE261983, HRA000395, GSE183852,
745 GSE167186, GSE180298, GSE243977, GSE136831, GSE255690, and GSE182786 for
746 a total of 85 samples from 9 tissues and 2 age groups (young and old) in human (Table
747 S3). Cell Ranger (v7.1.0) was used to process the raw data and generate UMI count
748 matrices.

749 The filtered UMI count matrices were imported into R using the Read10X function
750 and then analyzed using the R package Seurat (v4.9.9). After merging Seurat objects
751 for samples from the same tissue, cells with > 500 expressed genes and < 50% of

752 mitochondrial reads were retained. The NormalizeData function was used to normalize
753 the data. The identification of variable genes was conducted using the
754 FindVariableFeatures function, resulting in the selection of 5000 genes for subsequent
755 analysis. Principal component analysis (PCA) was performed using the RunPCA
756 function, and the RunHarmony function was used to remove batch effects from the data.
757 The top 20 PCs were used for clustering and U-MAP analysis. The FindAllMarkers
758 function was used to calculate the marker genes for each cluster. Cell types were
759 identified based on the expression levels of canonical marker genes (Table S4). Genes
760 expressed in each cell type (positive cell rate ≥ 0.01) were used for subsequent analysis.
761 We performed differential expression analysis between young and old groups using
762 the FindMarkers function in Seurat. Oocytes are not subjected to differential expression
763 analysis, as they are essentially absent from old ovaries. Genes with |Log2FC| ≥ 0.25
764 and adjusted P-values < 0.05 were defined as aging-related DEGs. Aging-related DEGs
765 in each category of lncRNA genes were analyzed for GO enrichment analysis. We used
766 the guilt-by-association method to associate the lncRNA genes with GO entries based
767 on their co-expressed protein-coding genes.

768 **Aging-related Differential Expression Analysis in Mouse**

769 We downloaded RNA-seq data from GSE225576 for a total of 96 samples from 8
770 tissues of 6-, 24-, and 30-months old male C57BL/6J mice. These RNA-seq data were
771 processed as before. Differential expression analysis was performed between 6- and 24-
772 or 30-month-old mice using DESeq2 (v.1.36.0) [84].

773 **Animals and Treatment**

774 The young (3-month-old) and aged (18-month-old) female C57BL/6L mice used in this
775 study were obtained from the Experimental Animal Center of Huazhong Agricultural
776 University. Animal husbandry and management strictly adhered to Good Laboratory
777 Practice standards. The animals were housed in an SPF-grade facility under controlled
778 environmental conditions, with the temperature maintained at 22 ± 1°C and a 12-hour
779 light/dark cycle. All mice were provided ad libitum access to sterilized feed and
780 drinking water. Animal care and experimental protocols were conducted in accordance

with the Guidelines for the Care and Use of Laboratory Animals issued by the National Research Council and approved by the Institutional Animal Care and Use Committee of Huazhong Agricultural University.

Total RNA Extraction and Quantitative Real-time PCR

Total RNA was isolated using TRIzol reagent (Catalog No. R401-01, Vazyme, Nanjing, China) following the manufacturer's protocol. cDNA synthesis was performed with HiScript II Q RT SuperMix for qPCR (+gDNA wiper) (Catalog No. R223-01, Vazyme, Nanjing, China) according to the manufacturer's instructions. RT-qPCR was conducted using ChamQ Universal SYBR qPCR Master Mix (Catalog No. Q711-02, Vazyme, Nanjing, China). Relative RNA expression was calculated using the Ct ($2^{-\Delta\Delta Ct}$) method. The expression of target genes was normalized to the expression of *GAPDH*. The RT-qPCR primers are listed in Table S7.

Ethical statement

The animal experiments were approved by the Animal Ethics Committees of Huazhong Agricultural University and of Institute of Hydrobiology, Chinese Academy of Sciences (approval ID: HZAUMO-2018-070, SYXK2015-0084). All the experiments were performed in accordance with the Good Practice Guidelines for the Use of Animal Research, Experiments, and Teaching.

Data availability

The raw sequence data reported in this paper have been deposited in the Genome Sequence Archive [85] in National Genomics Data Center [86], China National Center for Bioinformation / Beijing Institute of Genomics, Chinese Academy of Sciences (GSA: CRA023933, reviewer link: https://ngdc.cncb.ac.cn/gsa/s/QV6bw2mz; BioProject: PRJCA037421) that are publicly accessible at https://ngdc.cncb.ac.cn/gsa. The genomic annotations containing all pseudogene-associated lncRNA genes and lncRNA gene classification information used in this study have been deposited in the OMIX, China National Center for Bioinformation / Beijing Institute of Genomics,

Chinese Academy of Sciences (https://ngdc.cncb.ac.cn/omix: accession no. OMIX009515, reviewer link: https://ngdc.cncb.ac.cn/omix/preview/CNojc0bX).

Code availability

All scripts used in this manuscript have been deposited in Github (https://github.com/Z-H-Zhang/Pseudogene-associated_lncRNA_gene) and BioCode (https://ngdc.cncb.ac.cn/biocode: accession no. BT007886).

CRediT author statement

Ze-Hao Zhang: Investigation, Data curation, Formal analysis, Validation, Visualization, Writing – original draft, Writing – review & editing. **Bo-Han Li:** Investigation, Validation, Writing – review & editing. **Yin-Wei Wang:** Investigation, Visualization, Writing – review & editing. **Sheng Hu Qian:** Investigation, Visualization. **Lu Chen:** Methodology. **Meng-Wei Shi:** Software, Visualization. **Hao Zuo:** Investigation, Formal analysis, Visualization, Writing – review & editing, Supervision. **Zhen-Xia Chen:** Conceptualization, Resources, Writing – original draft, Writing – review & editing, Supervision, Funding acquisition. All authors have read and approved the final manuscript.

Competing interests

The authors declare no confict of interest.

Acknowledgements

This study was supported by National Key Research and Development Program of China (2023YFF1001000), The science and technology major program of Hubei Province (2021ABA011), the Foundation of Hubei Hongshan Laboratory (2021hszd012, 2022hszd024, 2022hszd028), and HZAU-AGIS Cooperation Fund (SZYJY2021010). We thank Fan Yao, Shuai Zhang and all members of Chen lab for helpful discussions.

ORCID

0000-0002-2190-491X (Ze-Hao Zhang)

0000-0002-1664-1393 (Bo-Han Li)
0000-0003-0474-9257 (Yin-Wei Wang)
0000-0002-6194-3876 (Sheng Hu Qian)
0000-0003-4031-5333 (Lu Chen)
0000-0001-6027-9742 (Meng-Wei Shi)
0009-0004-0031-6256 (Hao Zuo)
0000-0003-0474-902X (Zhen-Xia Chen)

References

[1] Sarropoulos I, Marin R, Cardoso-Moreira M, Kaessmann H. Developmental dynamics of lncRNAs across mammalian organs and species. Nature 2019;571:510−4.

[2] Herman AB, Tsitsipatis D, Gorospe M. Integrated lncRNA function upon genomic and epigenomic regulation. Mol Cell 2022;82:2252−66.

[3] Zhou B, Ji B, Liu K, Hu G, Wang F, Chen Q, et al. EVLncRNAs 2.0: an updated database of manually curated functional long non-coding RNAs validated by low-throughput experiments. Nucleic Acids Res 2021;49:D86−d91.

[4] Gao F, Cai Y, Kapranov P, Xu D. Reverse-genetics studies of lncRNAs-what we have learnt and paths forward. Genome Biol 2020;21:93.

[5] Han X, Luo S, Peng G, Lu JY, Cui G, Liu L, et al. Mouse knockout models reveal largely dispensable but context-dependent functions of lncRNAs during development. J Mol Cell Biol 2018;10:175−8.

[6] Liang W-W, Müller S, Hart SK, Wessels H-H, Méndez-Mancilla A, Sookdeo A, et al. Transcriptome-scale RNA-targeting CRISPR screens reveal essential lncRNAs in human cells. Cell 2024;187:7637−54.e29.

[7] Balasubramanian S, Zheng D, Liu YJ, Fang G, Frankish A, Carriero N, et al. Comparative analysis of processed ribosomal protein pseudogenes in four mammalian genomes. Genome Biol 2009;10:R2.

[8] Sisu C, Pei B, Leng J, Frankish A, Zhang Y, Balasubramanian S, et al. Comparative

865 analysis of pseudogenes across three phyla. PNAS 2014;111:13361−6.
866 [9] Chen X, Wan L, Wang W, Xi WJ, Yang AG, Wang T. Re-recognition of pseudogenes:
867 from molecular to clinical applications. Theranostics 2020;10:1479−99.
868 [10] Cheetham SW, Faulkner GJ, Dinger ME. Overcoming challenges and dogmas to
869 understand the functions of pseudogenes. Nat Rev Genet 2020;21:191−201.
870 [11] Zhang Z, Harrison PM, Liu Y, Gerstein M. Millions of years of evolution preserved:
871 a comprehensive catalog of the processed pseudogenes in the human genome.
872 Genome Res 2003;13:2541−58.
873 [12] Torrents D, Suyama M, Zdobnov E, Bork P. A genome-wide survey of human
874 pseudogenes. Genome Res 2003;13:2559−67.
875 [13] Harrison PM, Hegyi H, Balasubramanian S, Luscombe NM, Bertone P, Echols N,
876 et al. Molecular fossils in the human genome: identification and analysis of the
877 pseudogenes in chromosomes 21 and 22. Genome Res 2002;12:272−80.
878 [14] Qian SH, Chen L, Xiong Y-L, Chen Z-X. Evolution and function of
879 developmentally dynamic pseudogenes in mammals. Genome Biol 2022;23:235.
880 [15] Rapicavoli NA, Qu K, Zhang J, Mikhail M, Laberge R-M, Chang HY. A
881 mammalian pseudogene lncRNA at the interface of inflammation and anti-
882 inflammatory therapeutics. eLife 2013;2:e00762.
883 [16] Scarola M, Comisso E, Pascolo R, Chiaradia R, Maria Marion R, Schneider C, et
884 al. Epigenetic silencing of *Oct4* by a complex containing SUV39H1 and *Oct4*
885 pseudogene lncRNA. Nat Commun 2015;6:7631.
886 [17] Liu J, Liu Z-X, Wu Q-N, Lu Y-X, Wong C-W, Miao L, et al. Long noncoding RNA
887 *AGPG* regulates PFKFB3-mediated tumor glycolytic reprogramming. Nat
888 Commun 2020;11:1507.
889 [18] Feng J, Yang G, Liu Y, Gao Y, Zhao M, Bu Y, et al. LncRNA PCNAP1 modulates
890 hepatitis B virus replication and enhances tumor growth of liver cancer.
891 Theranostics 2019;9:5227−45.
892 [19] Lai Y, Li J, Zhong L, He X, Si X, Sun Y, et al. The pseudogene PTENP1 regulates
893 smooth muscle cells as a competing endogenous RNA. Clin Sci (Lond)

894 2019;133:1439−55.
895 [20] Zhang R, Kuo R, Coulter M, Calixto CPG, Entizne JC, Guo W, et al. A high-
896 resolution single-molecule sequencing-based *Arabidopsis* transcriptome using
897 novel methods of Iso-seq analysis. Genome Biol 2022;23:149.
898 [21] Wang K, Wang D, Zheng X, Qin A, Zhou J, Guo B, et al. Multi-strategic RNA-seq
899 analysis reveals a high-resolution transcriptional landscape in cotton. Nat
900 Commun 2019;10:4714.
901 [22] Zhang Z, Carriero N, Zheng D, Karro J, Harrison PM, Gerstein M. PseudoPipe: an
902 automated pseudogene identification pipeline. Bioinformatics 2006;22:1437−9.
903 [23] Sisu C, Muir P, Frankish A, Fiddes I, Diekhans M, Thybert D, et al. Transcriptional
904 activity and strain-specific history of mouse pseudogenes. Nat Commun
905 2020;11:3695.
906 [24] Lam HYK, Khurana E, Fang G, Cayting P, Carriero N, Cheung K-H, et al.
907 Pseudofam: the pseudogene families database. Nucleic Acids Res
908 2008;37:D738−D43.
909 [25] Hillier LW, Miller W, Birney E, Warren W, Hardison RC, Ponting CP, et al.
910 Sequence and comparative analysis of the chicken genome provide unique
911 perspectives on vertebrate evolution. Nature 2004;432:695−716.
912 [26] Lee S-H, Eldi P, Cho S-Y, Rangasamy D. Control of chicken CR1 retrotransposons
913 is independent of Dicer-mediated RNA interference pathway. BMC Biol
914 2009;7:53.
915 [27] Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, et al. The
916 zebrafish reference genome sequence and its relationship to the human genome.
917 Nature 2013;496:498−503.
918 [28] Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, et al. The
919 evolution of lncRNA repertoires and expression patterns in tetrapods. Nature
920 2014;505:635−40.
921 [29] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP:
922 the NCBI database of genetic variation. Nucleic Acids Res 2001;29:308−11.

[30] Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 2014;46:310−5.

[31] Cai D, Han JJ. Aging-associated lncRNAs are evolutionarily conserved and participate in NFκB signaling. Nat Aging 2021;1:438−53.

[32] Guttman M, Rinn JL. Modular regulatory principles of large non-coding RNAs. Nature 2012;482:339−46.

[33] Jiang S, Cheng SJ, Ren LC, Wang Q, Kang YJ, Ding Y, et al. An expanded landscape of human long noncoding RNA. Nucleic Acids Res 2019;47:7842−56.

[34] de Goede OM, Nachun DC, Ferraro NM, Gloudemans MJ, Rao AS, Smail C, et al. Population-scale tissue transcriptomics maps long non-coding RNAs to complex disease. Cell 2021;184:2633−48.e19.

[35] Schuettengruber B, Martinez A-M, Iovino N, Cavalli G. Trithorax group proteins: switching genes on and keeping them active. Nat Rev Mol Cell Biol 2011;12:799−814.

[36] Wang M, Chen Z, Zhang Y. CBP/p300 and HDAC activities regulate H3K27 acetylation dynamics and zygotic genome activation in mouse preimplantation embryos. EMBO J 2022;41:e112012.

[37] Jin Q, Yu LR, Wang L, Zhang Z, Kasper LH, Lee JE, et al. Distinct roles of GCN5/PCAF-mediated H3K9ac and CBP/p300-mediated H3K18/27ac in nuclear receptor transactivation. EMBO J 2011;30:249−62.

[38] Ernst J, Kellis M. Chromatin-state discovery and genome annotation with ChromHMM. Nat Protoc 2017;12:2478−92.

[39] Statello L, Guo C-J, Chen L-L, Huarte M. Gene regulation by long non-coding RNAs and its biological functions. Nat Rev Mol Cell Biol 2021;22:96−118.

[40] Hentze MW, Castello A, Schwarzl T, Preiss T. A brave new world of RNA-binding proteins. Nat Rev Mol Cell Biol 2018;19:327−41.

[41] Huang W, Xiong T, Zhao Y, Heng J, Han G, Wang P, et al. Computational prediction and experimental validation identify functionally conserved lncRNAs

952 from zebrafish to human. Nat Genet 2024;56:124−35.
953 [42] Lee S, Kopp F, Chang T-C, Sataluri A, Chen B, Sivakumar S, et al. Noncoding
954 RNA *NORAD* regulates genomic stability by sequestering PUMILIO Proteins.
955 Cell 2016;164:69−80.
956 [43] Hosono Y, Niknafs YS, Prensner JR, Iyer MK, Dhanasekaran SM, Mehra R, et al.
957 Oncogenic role of *THOR*, a conserved cancer/testis long non-coding RNA. Cell
958 2017;171:1559−72.e20.
959 [44] Wu G, Su J, Zeng L, Deng S, Huang X, Ye Y, et al. LncRNA *BCAN-AS1* stabilizes
960 c-Myc via N6-methyladenosine-mediated binding with SNIP1 to promote
961 pancreatic cancer. Cell Death Differ 2023;30:2213−30.
962 [45] Liu Y, Yang D, Liu T, Chen J, Yu J, Yi P. N6-methyladenosine-mediated gene
963 regulation and therapeutic implications. Trends Mol Med 2023;29:454−67.
964 [46] Gong Y, Liang Y, Liu J, Wei J, Zhang S, Chen F, et al. DDX24 is essential for cell
965 cycle regulation in vascular smooth muscle cells during vascular development via
966 binding to *FANCA* mRNA. Arterioscler, Thromb, Vasc Biol 2023;43:1653−67.
967 [47] Randolph ME, Afifi M, Gorthi A, Weil R, Wilky BA, Weinreb J, et al. RNA
968 helicase DDX3 regulates RAD51 localization and DNA damage repair in Ewing
969 sarcoma. iScience 2024;27:108925.
970 [48] Riva V, Garbelli A, Casiraghi F, Arena F, Trivisani Claudia I, Gagliardi A, et al.
971 Novel alternative ribonucleotide excision repair pathways in human cells by
972 DDX3X and specialized DNA polymerases. Nucleic Acids Res
973 2020;48:11551−65.
974 [49] Gu J, Chen J, Yin Q, Dong M, Zhang Y, Chen M, et al. lncRNA JPX-enriched
975 chromatin microenvironment mediates vascular smooth muscle cell senescence
976 and promotes atherosclerosis. Arterioscler, Thromb, Vasc Biol 2024;44:156−76.
977 [50] Ruvkun G, Lehrbach N. Regulation and functions of the ER-associated Nrf1
978 transcription factor. Cold Spring Harb Perspect Biol 2023;15.
979 [51] Suryadevara V, Hudgins AD, Rajesh A, Pappalardo A, Karpova A, Dey AK, et al.
980 SenNet recommendations for detecting senescent cells in different tissues. Nat Rev

981 Mol Cell Biol 2024.
982 [52] Vernier M, Bourdeau V, Gaumont-Leclerc MF, Moiseeva O, Bégin V, Saad F, et al.
983 Regulation of E2Fs and senescence by PML nuclear bodies. Genes Dev
984 2011;25:41−50.
985 [53] Yang S, Liu C, Jiang M, Liu X, Geng L, Zhang Y, et al. A single-nucleus
986 transcriptomic atlas of primate liver aging uncovers the pro-senescence role of
987 SREBP2 in hepatocytes. Protein Cell 2023;15:98−120.
988 [54] McIntyre RL, Liu YJ, Hu M, Morris BJ, Willcox BJ, Donlon TA, et al.
989 Pharmaceutical and nutraceutical activation of FOXO3 for healthy longevity.
990 Ageing Res Rev 2022;78:101621.
991 [55] Liu G, Chen H, Liu H, Zhang W, Zhou J. Emerging roles of SIRT6 in human
992 diseases and its modulators. Med Res Rev 2021;41:1089−137.
993 [56] Jing Y, Jiang X, Ji Q, Wu Z, Wang W, Liu Z, et al. Genome-wide CRISPR
994 activation screening in senescent cells reveals SOX5 as a driver and therapeutic
995 target of rejuvenation. Cell Stem Cell 2023;30:1452−71.e10.
996 [57] Johnson SC, Rabinovitch PS, Kaeberlein M. mTOR is a key modulator of ageing
997 and age-related disease. Nature 2013;493:338−45.
998 [58] Bou Sleiman M, Jha P, Houtkooper R, Williams RW, Wang X, Auwerx J. The gene-
999 regulatory footprint of aging highlights conserved central regulators. Cell Rep
1000 2020;32:108203.
1001 [59] Tang H, Xiong Q, Yin M, Feng H, Yao F, Xiao X, et al. LncRNA PVT1 delays skin
1002 photoaging by sequestering miR-551b-3p to release AQP3 expression via ceRNA
1003 mechanism. Apoptosis 2023;28:912−24.
1004 [60] Du H, Zhang H, Yang R, Qiao L, Shao H, Zhang X. Small interfering RNA-
1005 induced silencing lncRNA PVT1 inhibits atherosclerosis via inactivating the
1006 MAPK/NF-κB pathway. Aging 2021;13:24449−63.
1007 [61] Palazzo AF, Koonin EV. Functional long non-coding RNAs evolve from junk
1008 transcripts. Cell 2020;183:1151−61.
1009 [62] Ma Y, Liu S, Gao J, Chen C, Zhang X, Yuan H, et al. Genome-wide analysis of

1010 pseudogenes reveals HBBP1's human-specific essentiality in erythropoiesis and
1011 implication in β-thalassemia. Dev Cell 2021;56:478−93.e11.
1012 [63] Li Z, Liu L, Feng C, Qin Y, Xiao J, Zhang Z, et al. LncBook 2.0: integrating human
1013 long non-coding RNAs with multi-omics annotations. Nucleic Acids Res
1014 2022;51:D186−D91.
1015 [64] Kaessmann H. Origins, evolution, and phenotypic impact of new genes. Genome
1016 Res 2010;20:1313−26.
1017 [65] Lister NC, Johnsson P, Waters PD, Morris KV. Pseudogenes: a novel source of
1018 trans-acting antisense RNAs. Methods Mol Biol 2014;1167:213−26.
1019 [66] Milligan MJ, Lipovich L. Pseudogene-derived lncRNAs: emerging regulators of
1020 gene expression. Front Genet 2015;5.
1021 [67] Salmela L, Rivals E. LoRDEC: accurate and efficient long read error correction.
1022 Bioinformatics 2014;30:3506−14.
1023 [68] Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics
1024 2018;34:3094−100.
1025 [69] Kang YJ, Yang DC, Kong L, Hou M, Meng YQ, Wei L, et al. CPC2: a fast and
1026 accurate coding potential calculator based on sequence intrinsic features. Nucleic
1027 Acids Res 2017;45:W12−w6.
1028 [70] Sun L, Luo H, Bu D, Zhao G, Yu K, Zhang C, et al. Utilizing sequence intrinsic
1029 composition to classify protein-coding and long non-coding transcripts. Nucleic
1030 Acids Res 2013;41:e166.
1031 [71] Li A, Zhang J, Zhou Z. PLEK: a tool for predicting long non-coding RNAs and
1032 messenger RNAs based on an improved k-mer scheme. BMC Bioinf 2014;15:311.
1033 [72] Wang G, Yin H, Li B, Yu C, Wang F, Xu X, et al. Characterization and identification
1034 of long non-coding RNAs based on feature relationship. Bioinformatics
1035 2019;35:2949−56.
1036 [73] Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, et al.
1037 InterProScan: protein domains identifier. Nucleic Acids Res 2005;33:W116−20.
1038 [74] Guo C, Duan Y, Ye W, Zhang W, Cheng Y, Shi M, et al. FishGET: a fish gene

1039 expression and transcriptome database with improved accuracy and visualization.
1040 iScience 2023;26:106539.
1041 [75] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR:
1042 ultrafast universal RNA-seq aligner. Bioinformatics 2013;29:15−21.
1043 [76] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence
1044 alignment/map format and SAMtools. Bioinformatics 2009;25:2078−9.
1045 [77] Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for
1046 assigning sequence reads to genomic features. Bioinformatics 2013;30:923−30.
1047 [78] Qiuchen M, Xinze W, Chen L, Jiaqi L, Xi X, Sijie C, et al. The full set of potential
1048 open regions (PORs) in the human genome defined by consensus peaks of ATAC-
1049 seq data. bioRxiv 2023:2023.05.30.542889.
1050 [79] Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, et al.
1051 deepTools2: a next generation web server for deep-sequencing data analysis.
1052 Nucleic Acids Res 2016;44:W160−5.
1053 [80] Yevshin I, Sharipov R, Kolmykov S, Kondrakhin Y, Kolpakov F. GTRD: a
1054 database on gene transcription regulation-2019 update. Nucleic Acids Res
1055 2019;47:D100−d5.
1056 [81] Liu S, Zhu A, He C, Chen M. REPIC: a database for exploring the N6-
1057 methyladenosine methylome. Genome Biol 2020;21:100.
1058 [82] Zhao W, Zhang S, Zhu Y, Xi X, Bao P, Ma Z, et al. POSTAR3: an updated platform
1059 for exploring post-transcriptional regulation coordinated by RNA-binding proteins.
1060 Nucleic Acids Res 2022;50:D287−d94.
1061 [83] Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: a universal
1062 enrichment tool for interpreting omics data. Innovation (Camb) 2021;2:100141.
1063 [84] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion
1064 for RNA-seq data with DESeq2. Genome Biol 2014;15:550.
1065 [85] Chen T, Chen X, Zhang S, Zhu J, Tang B, Wang A, et al. The Genome Sequence
1066 Archive Family: toward explosive data growth and diverse data types. Genomics
1067 Proteomics Bioinformatics 2021.

[86] CNCB-NGDC Members and Partners. Database resources of the National Genomics Data Center, China National Center for Bioinformation in 2024. Nucleic Acids Res 2024;52:D18−d32.

Figure legends

Figure 1 Identification of pseudogene-associated lncRNA genes

A. Tissue samples from zebrafish, chicken and mouse. Somatic tissues were pooled into one sample for Iso-seq library preparation for each sex. **B**. Length of transcripts in Ensembl annotation and PacBio Iso-seq. **C**. Exon number of transcripts in Ensembl annotation and PacBio Iso-seq. **D**. Number of lncRNA annotations in each species. **E**. Number of pseudogenes in each species. **F**. Methods to identify pseudogene-associated lncRNA genes. **G**. Number of pseudogene-associated lncRNA genes in each species. Bar plot showing the proportion of two categories among pseudogene-associated lncRNA genes. **H**. Expression patterns of zebrafish novel lncRNA gene *PB.7298* and mouse novel lncRNA gene *PB.5227* which are detected by Iso-seq and are identified as pseudogene-associated lncRNA genes. Iso-seq, isoform sequencing.

Figure 2 lncRNA genes are enriched at pseudogenes

A. Enrichment of sense and antisense lncRNA genes at pseudogenes in each species. $*P \leq 0.05$, $**P \leq 0.01$, $***P \leq 0.001$. **B**. Sequence conservation (PhastCons score), for overlapping and non-overlapping regions within the pseudogenes overlapping with sense and antisense lncRNA genes in human and mouse. **C**. Genomic profile depicts a high conservation of regions within the pseudogene HOS_PG_6183 overlapping the *ROCK1P1* exons. The per-base phastCons, and the phyloP conservation scores are presented.

Figure 3 Functionality characterization and expression profiles of pseudogene-associated lncRNA genes

A. Distribution of PAS lncRNA genes, PAA lncRNA genes, and NPA lncRNA genes in

1097 the branches of the human phylogenetic tree. Tree tips, lncRNA proportions for each
1098 age group. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, two-sided Fisher's exact test. **B**.
1099 Enrichment of PAS lncRNA genes, PAA lncRNA genes, and NPA lncRNA genes at
1100 different evolutionary ages in human. A total of 2000 randomly selected NPA lncRNA
1101 genes were utilized as controls. **C**. Exonic sequence conservation (PhastCons score),
1102 for CDS, UTR, PAS lncRNA genes, PAA lncRNA genes, NPA lncRNA genes, and
1103 random intergenic regions in human. **D**. Exon common SNP density of CDS, UTR, PAS
1104 lncRNA genes, PAA lncRNA genes, NPA lncRNA genes, and random intergenic regions
1105 in human. **E**. Transcript lengths of protein-coding genes, PAS lncRNA genes, PAA
1106 lncRNA genes, and NPA lncRNA genes in human. **F**. Exonic sequence conservation
1107 (PhastCons score), for PAS lncRNA genes, PAA lncRNA genes, and NPA lncRNA
1108 genes in each age group in human. **G**. Maximum expression levels of protein-coding
1109 genes, PAS lncRNA genes, PAA lncRNA genes, and NPA lncRNA genes in human. **H**.
1110 Expression proportions of protein-coding genes, PAS lncRNA genes, PAA lncRNA
1111 genes, and NPA lncRNA genes at a series of expression thresholds in human. **I**. Mean
1112 expression proportions of protein-coding genes, PAS lncRNA genes, PAA lncRNA
1113 genes, and NPA lncRNA genes in various tissues in human. Error bars indicate standard
1114 deviation of the proportion of expression. Legend is the same as in Figure 3H. PAS,
1115 pseudogene-associated sense; PAA, pseudogene-associated antisense; NPA, non-
1116 pseudogene-associated; Ma, million years ago; CDS, coding sequence; UTR,
1117 untranslated region; SNP, single nucleotide polymorphism.
1118
1119 **Figure 4 Epigenomic features of pseudogene-associated lncRNA genes**
1120 **A**. Distribution of housekeeping consensus peaks from ATAC-seq in a total of 4kb
1121 regions upstream and downstream of protein-coding gene, PAS lncRNA genes, PAA
1122 lncRNA genes, and NPA lncRNA genes TSSs, and random intergenic regions. **B**.
1123 Distribution of DHS in a total of 4kb regions upstream and downstream of protein-
1124 coding gene, PAS lncRNA genes, PAA lncRNA genes, and NPA lncRNA genes TSSs,
1125 and random intergenic regions. **C**. Distribution of H3K4me3, H3K4me2, H3K9ac, and

1126 H3K27ac peaks in a total of 4kb regions upstream and downstream of protein-coding
1127 gene, PAS lncRNA genes, PAA lncRNA genes, and NPA lncRNA genes TSSs, and
1128 random intergenic regions. **D**. Distribution of ten chromatin modifiers binding sites in
1129 a total of 2kb regions upstream and downstream of protein-coding gene, PAS lncRNA
1130 genes, PAA lncRNA genes, and NPA lncRNA genes TSSs, and random intergenic
1131 regions. **E**. DNA methylation in each 200bp window within a total of 4kb regions
1132 upstream and downstream of protein-coding gene, PAS lncRNA genes, PAA lncRNA
1133 genes, and NPA lncRNA genes TSSs, and random intergenic regions. Dotted lines
1134 indicate average DNA methylation values over the entire region. Legend is the same as
1135 in Figure 4A. **F**. Ratios of active promoter chromHMM state regions in the proximal
1136 promoters of protein-coding genes, PAS lncRNA genes, PAA lncRNA genes, NPA
1137 lncRNA genes, and random intergenic regions. TSS, transcription start site; DHS,
1138 DNase I hypersensitivity site.
1139

1140 **Figure 5 RNA-protein binding related features and putative functions of**
1141 **pseudogene-associated lncRNA genes**

1142 **A**. Density of m6A sites in protein-coding gene transcripts, PAS lncRNAs, PAA
1143 lncRNAs, and NPA lncRNAs. **B**. Richness of RBPs bound to protein-coding gene
1144 transcripts, PAS lncRNAs, PAA lncRNAs, and NPA lncRNAs. **C**. Bray-Curtis
1145 similarity between RBPs bound to pseudogene-associated lncRNAs and RBPs bound
1146 to transcripts of parent genes of pseudogenes. Two types of gene pairs were utilized as
1147 controls: pseudogene-associated lncRNA genes and random protein-coding genes that
1148 exhibited a matched expression level with parent genes of pseudogenes, as well as
1149 parent genes of pseudogenes and random lncRNA genes that exhibited a matched
1150 expression level with pseudogene-associated lncRNA genes. **D**. Top 10 RBPs enriched
1151 for PAS and PAA lncRNAs. A total of 2000 randomly selected NPA lncRNAs were used
1152 as controls. **E**. GO enrichment analysis to annotate the function of PAS lncRNA genes
1153 in human. The representative GO biological processes are shown. m6A, N6-
1154 methyladenosine; RBP, RNA-binding protein; GO, Gene Ontology.

1155

1156 **Figure 6 Aging-related expression changes of pseudogene-associated lncRNA**
1157 **genes across nine human tissues**

1158 **A**. Proportions of aging-related DEGs for protein-coding genes, PAS lncRNA genes,
1159 PAA lncRNA genes, and NPA lncRNA genes in each tissue. **B**. Odds ratios and
1160 confidence intervals (95%) between the proportions of aging-related DEGs in each
1161 category of lncRNA genes in nine tissues. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, two-
1162 sided Fisher's exact tests were performed for comparisons between each category of
1163 lncRNA genes. **C**. Among aging-related DEGs, proportions of tissue-shared aging-
1164 related DEGs for protein-coding genes, PAS lncRNA genes, PAA lncRNA genes, and
1165 NPA lncRNA genes. **D**. Number of tissues in which the genes exhibited aging-related
1166 differential expression for protein-coding genes, PAS lncRNA genes, PAA lncRNA
1167 genes, and NPA lncRNA genes in each expression breadth group. The diamond dots
1168 represent the average. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, two-sided Wilcoxon rank
1169 sum test. **E**. Aging-related differential expression of *JPX* in each cell type of each tissue.
1170 **F**. GO enrichment analysis based on guilt-by-association method to annotate the
1171 function of aging-related DEGs in each category of lncRNA genes. The representative
1172 GO biological processes are shown. A total of 2000 randomly selected NPA lncRNA
1173 genes were used as controls for comparison. DEG, differentially expressed gene; Mic,
1174 microglia; FB, fibroblast; Myeloid, myeloid cell; HSC, hematopoietic stem cell; LMPP,
1175 lympho-myeloid primed progenitor; GMP, granulocyte-monocyte progenitor; MONO,
1176 monocyte; pDC, plasmacytoid dendritic cell; CLP, common lymphoid progenitor; MEP,
1177 megakaryocyte-erythroid progenitor; Mk, megakaryocyte; Ery, erythroid; Basophil,
1178 basophil cell; Prolif, proliferative cell; Macro, macrophage; GC, granulosa cell; T&S,
1179 theca and stroma cell; SMC, smooth muscle cell; EC, endothelial cell; Late SPC, late
1180 primary spermatocyte.

1181

1182 **Figure 7 Aging-related expression changes of pseudogene-associated lncRNA**
1183 **genes in mouse**

1184 **A**. Aging-related expression changes of the top 3 PAS and PAA lncRNA genes
1185 differentially expressed in the most mouse tissues. Statistical significance labels are
1186 based on adjusted p-values from DESeq2, *$P \leq 0.05$, **$P \leq 0.01$, ***$P \leq 0.001$. **B**.
1187 Expression levels of *D17H6S56E-5* in five tissues from 3 month and 18 month old mice,
1188 revealed by RT-qPCR. Expression levels are shown as means. Error bars indicate the
1189 standard deviation of expression levels. 3 month, n = 3; 18 month, n = 3 mice. **$P <$
1190 0.01, ***$P < 0.001$, two-sided Student's T test. **C**. Expression levels of *Gm36161* in
1191 five tissues from 3 month and 18 month old mice, revealed by RT-qPCR. Expression
1192 levels are shown as means. Error bars indicate the standard deviation of expression
1193 levels. 3 month, n = 3; 18 month, n = 3 mice. **$P < 0.01$, ***$P < 0.001$, two-sided
1194 Student's T test. **D**. Orthologous pseudogene-associated lncRNA gene pairs that
1195 exhibited aging-related differential expression in both human and mouse. **E**. Expression
1196 levels of *Mou_XLOC_028665*, *Mou_XLOC_013922*, *Pvt1*, and *Mou_XLOC_000120*
1197 were quantified by RT-qPCR in five tissues from 3 month and 18 month old mice.
1198 Expression levels are shown as means. Error bars indicate the standard deviation of
1199 expression levels. 3 month, n = 3; 18 month, n = 3 mice. *$P < 0.05$, **$P < 0.01$, ***P
1200 < 0.001, two-sided Student's T test. RT-qPCR, quantitative real-time PCR.
1201
1202

1203 ## Supplementary material

1204 **Figure S1 Pipeline for identification of pseudogenes, integration of lncRNA**
1205 **annotations, and identification of the pseudogene-associated lncRNA gene**
1206

1207 **Figure S2 Features of lncRNA genes, pseudogenes, and pseudogene-associated**
1208 **lncRNA genes**

1209 **A**. Transcript lengths of Ensemble lncRNA genes, novel lncRNA genes annotated by
1210 long-read transcriptome sequencing, and novel lncRNA genes annotated by short-read
1211 transcriptome sequencing. **B**. Exonic sequence conservation for Ensemble lncRNA
1212 genes, novel lncRNA genes annotated by long-read transcriptome sequencing, and

1213 novel lncRNA genes annotated by short-read transcriptome sequencing. **C**. Maximum
1214 expression levels of Ensemble lncRNA genes, novel lncRNA genes annotated by long-
1215 read transcriptome sequencing, and novel lncRNA genes annotated by short-read
1216 transcriptome sequencing. **D**. Proportion of the three categories of pseudogenes in each
1217 species. **E**. Desnsity of the three categories of pseudogenes in each species. **F**. Pearson
1218 correlation analysis between the number of pseudogene-associated lncRNA genes and
1219 the total number of lncRNA genes. **G**. Bar plot showing the proportion of novel lncRNA
1220 genes among pseudogene-associated lncRNA genes.
1221
1222 **Figure S3 Enrichment of sense and antisense lncRNA genes at each category of**
1223 **pseudogenes in each species**
1224
1225 **Figure S4 Functionality characterization of pseudogene-associated lncRNA**
1226 **genes in human and mouse**
1227 **A**. Distribution of PAS lncRNA genes, PAA lncRNA genes, and NPA lncRNA genes in
1228 the branches of the mouse phylogenetic tree. Tree tips, lncRNA proportions for each
1229 age group. $*P < 0.05$, $**P < 0.01$, $***P < 0.001$, two-sided Fisher's exact test. **B**.
1230 Enrichment of PAS lncRNA genes, PAA lncRNA genes, and NPA lncRNA genes at
1231 different evolutionary ages in mouse. A total of 2000 randomly selected NPA lncRNAs
1232 were utilized as controls. **C**. Exonic sequence conservation (PhastCons score), for CDS,
1233 UTR, PAS lncRNA genes, PAA lncRNA genes, NPA lncRNA genes, and random
1234 intergenic regions in mouse. **D**. Exonic sequence conservation (PhastCons score), for
1235 PAS lncRNA genes, PAA lncRNA genes, and NPA lncRNA genes in each age group in
1236 mouse. **E**. Mean derived allele frequency of non-CpG SNPs segregating in African
1237 populations (1000 Genomes Project). Error bars, 95% confidence intervals based on
1238 1000 bootstrap resampling replicates. **F**. CADD scores for SNPs in CDS, UTR, PAS
1239 lncRNA genes, PAA lncRNA genes, NPA lncRNA genes, and random intergenic regions
1240 in human. CADD, Combined Annotation-Dependent Depletion. **G**. Transcript lengths
1241 of protein-coding genes, PAS lncRNA genes, PAA lncRNA genes, and NPA lncRNA

genes in mouse. **H**. Average number of exons in transcripts of protein-coding genes, PAS lncRNA genes, PAA lncRNA genes, and NPA lncRNA genes in human and mouse. Error bars, 95% confidence intervals based on 1000 bootstrap resampling replicates.

Figure S5 Maximum expression levels of protein-coding genes, PAS lncRNA genes, PAA lncRNA genes, and NPA lncRNA genes in various tissues in human.

Figure S6 Expression profiles of pseudogene-associated lncRNA genes in human and mouse

A. Maximum expression levels of protein-coding genes, PAS lncRNA genes, PAA lncRNA genes, and NPA lncRNA genes in mouse. **B**. Expression proportions of protein-coding genes, PAS lncRNA genes, PAA lncRNA genes, and NPA lncRNA genes at a series of expression thresholds in mouse. **C**. Mean expression proportions of protein-coding genes, PAS lncRNA genes, PAA lncRNA genes, and NPA lncRNA genes in various tissues in mouse. Error bars indicate standard deviation of the proportion of expression. **D**. Maximum expression levels of protein-coding genes, PAS lncRNA genes, PAA lncRNA genes, and NPA lncRNA genes in various tissues in mouse. **E**. Number of tissues with expression of protein-coding genes, PAS lncRNA genes, PAA lncRNA genes, and NPA lncRNA genes in human and mouse.

Figure S7 RBP-binding capabilities and putative functions for each category of lncRNA genes

A. Shannon diversity index of RBPs bound to protein-coding gene transcripts, PAS lncRNAs, PAA lncRNAs, and NPA lncRNAs in human. **B**. Density of m6A sites in protein-coding gene transcripts, PAS lncRNAs, PAA lncRNAs, and NPA lncRNAs in each expression level in human. **C**. Richness of RBPs bound to protein-coding gene transcripts, PAS lncRNAs, PAA lncRNAs, and NPA lncRNAs in each expression level in human. **D**. Shannon diversity index of RBPs bound to protein-coding gene transcripts, PAS lncRNAs, PAA lncRNAs, and NPA lncRNAs in each expression level in human.

1271 **E**. Number of RBPs enriched by PAS lncRNAs and PAA lncRNAs. **F**. Number of co-
1272 expressed genes enriched by each category of lncRNA genes at each p-value level in
1273 human. A total of 2000 randomly selected NPA lncRNA genes were utilized as controls.
1274 **G**. Number of co-expressed genes enriched by each category of lncRNA genes at each
1275 p-value level in mouse. A total of 2000 randomly selected NPA lncRNA genes were
1276 utilized as controls. **H**. GO enrichment analysis to annotate the function of NPA
1277 lncRNA genes in mouse. **I**. GO enrichment analysis to annotate the function of PAS
1278 lncRNA genes in mouse. **J**. GO enrichment analysis to annotate the function of PAA
1279 lncRNA genes in mouse. **K**. GO enrichment analysis to annotate the function of PAA
1280 lncRNA genes in human.
1281
1282 **Figure S8 UMAP plots showing the annotated cell types in each tissue**
1283 Cell types are indicated by the different colors.
1284
1285 **Figure S9 Expression proportions for protein-coding genes, PAS lncRNA genes,**
1286 **PAA lncRNA genes, and NPA lncRNA genes in each cell type of each tissue**
1287
1288 **Figure S10 Expression proportions of protein-coding genes, PAS lncRNA genes,**
1289 **PAA lncRNA genes, and NPA lncRNA genes in each tissue**
1290
1291 **Figure S11 Number of cell types with expression of protein-coding genes, PAS**
1292 **lncRNA genes, PAA lncRNA genes, and NPA lncRNA genes**
1293
1294 **Figure S12 Proportions of up-regulated and down-regulated genes for protein-**
1295 **coding genes, PAS lncRNA genes, PAA lncRNA genes, and NPA lncRNA genes in**
1296 **each cell type of each tissue**
1297 Two-sided Fisher's exact tests were conducted to compare pseudogene-associated
1298 lncRNA genes with NPA lncRNA genes. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$.
1299

1300 **Figure S13 Among aging-related DEGs, proportions of tissue-specific genes**
1301 **(Expressed in only one tissue) for protein-coding genes, PAS lncRNA genes, PAA**
1302 **lncRNA genes, and NPA lncRNA genes**
1303 *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, two-sided Fisher's exact test.
1304
1305 **Figure S14 Enrichment of TF binding sites for each category of lncRNA genes in**
1306 **human**
1307 The representative TFs are shown. A total of 2000 randomly selected NPA lncRNA
1308 genes were utilized as controls. TF, transcription factor.
1309
1310 **Figure S15 Aging-related expression changes of pseudogene-associated lncRNA**
1311 **genes across eight mouse tissues**
1312 **A**. Proportions of aging-related DEGs for protein-coding genes, PAS lncRNA genes,
1313 PAA lncRNA genes, and NPA lncRNA genes in each tissue. **B**. Odds ratios and
1314 confidence intervals (95%) between the proportions of aging-related DEGs in each
1315 category of lncRNA genes in eight tissues. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, two-
1316 sided Fisher's exact tests were performed for comparisons between each category of
1317 lncRNA genes. **C**. Among aging-related DEGs, proportions of tissue-shared aging-
1318 related DEGs for protein-coding genes, PAS lncRNA genes, PAA lncRNA genes, and
1319 NPA lncRNA genes. **D**. Number of tissues in which the genes exhibited aging-related
1320 differential expression for protein-coding genes, PAS lncRNA genes, PAA lncRNA
1321 genes, and NPA lncRNA genes in each expression breadth group. The diamond dots
1322 represent the average. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, two-sided Wilcoxon rank
1323 sum test.
1324
1325 **Figure S16 Expression levels of *ENSMUSG00000121500*,**
1326 ***ENSMUSG00000121378*, *Mou_XLOC_017668*, and *Mou_XLOC_000586* in five**
1327 **tissues from 3 month and 18 month old mice, revealed by RT-qPCR**
1328 Expression levels are shown as means. Error bars indicate standard deviation of the

1329 expression levels. 3 month, n = 3; 18 month, n = 3 mice. *$P < 0.05$, **$P < 0.01$,
1330 ***$P < 0.001$, two-sided Student's T test.
1331
1332 **Figure S17 Aging-related expression changes for *Mou_XLOC_028665*,**
1333 ***Mou_XLOC_013922*, *Pvt1*, and *Mou_XLOC_000120*, revealed by RNA-seq**
1334 **analysis**
1335 Statistical significance labels are based on adjusted p-values from DESeq2, *$P \leq 0.05$,
1336 **$P \leq 0.01$, ***$P \leq 0.001$.
1337
1338 **Figure S18 Venn plot shows the overlap between LncBook database and our**
1339 **human lncRNA gene annotation**
1340 The pie plot on the right shows the sources of lncRNA genes that are only present in
1341 our human lncRNA gene annotation.
1342
1343 **Figure S19 Functionality characterization of PAS lncRNA genes at a series of**
1344 **overlap length thresholds in human**
1345 **A**. Enrichment of evolutionary ages. **B**. Exonic sequence conservation (PhastCons
1346 score). **C**. Transcript lengths. **D**. Maximum expression levels. **E**. Density of m6A sites.
1347 **F**. RBP richness. **G**. RBP diversity. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, Wilcoxon
1348 rank sum test.
1349
1350 **Figure S20 Functionality characterization of PAA lncRNA genes at a series of**
1351 **overlap length thresholds in human**
1352 **A**. Enrichment of evolutionary ages. **B**. Exonic sequence conservation (PhastCons
1353 score). **C**. Transcript lengths. **D**. Maximum expression levels. **E**. Density of m6A sites.
1354 **F**. RBP richness. **G**. RBP diversity. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, Wilcoxon
1355 rank sum test.
1356
1357 **Figure S21 Functionality characterization of PAS lncRNA genes at a series of**

1358 **overlap length thresholds in mouse**

1359 **A**. Enrichment of evolutionary ages. **B**. Exonic sequence conservation (PhastCons
1360 score). **C**. Transcript lengths. **D**. Maximum expression levels. *$P < 0.05$, **$P < 0.01$,
1361 ***$P < 0.001$, Wilcoxon rank sum test.

1362

1363 **Figure S22 Functionality characterization of PAA lncRNA genes at a series of**
1364 **overlap length thresholds in mouse**

1365 **A**. Enrichment of evolutionary ages. **B**. Exonic sequence conservation (PhastCons
1366 score). **C**. Transcript lengths. **D**. Maximum expression levels. *$P < 0.05$, **$P < 0.01$,
1367 ***$P < 0.001$, Wilcoxon rank sum test.

1368

1369 **Figure S23 Functionality characterization of PAS lncRNA genes at a series of**
1370 **expression level thresholds in human**

1371 **A**. Enrichment of evolutionary ages. **B**. Exonic sequence conservation (PhastCons
1372 score). **C**. Transcript lengths.

1373

1374 **Figure S24 Functionality characterization of PAA lncRNA genes at a series of**
1375 **expression level thresholds in human**

1376 **A**. Enrichment of evolutionary ages. **B**. Exonic sequence conservation (PhastCons
1377 score). **C**. Transcript lengths.

1378

1379 **Figure S25 Functionality characterization of PAS lncRNA genes at a series of**
1380 **expression level thresholds in mouse**

1381 **A**. Enrichment of evolutionary ages. **B**. Exonic sequence conservation (PhastCons
1382 score). **C**. Transcript lengths.

1383

1384 **Figure S26 Functionality characterization of PAA lncRNA genes at a series of**
1385 **expression level thresholds in mouse**

1386 **A**. Enrichment of evolutionary ages. **B**. Exonic sequence conservation (PhastCons

score). **C**. Transcript lengths.

Table S1 Pseudogene-associated lncRNA genes in eight vertebrates

Table S2 Functionality characterization of pseudogene-associated lncRNA genes in human and mouse

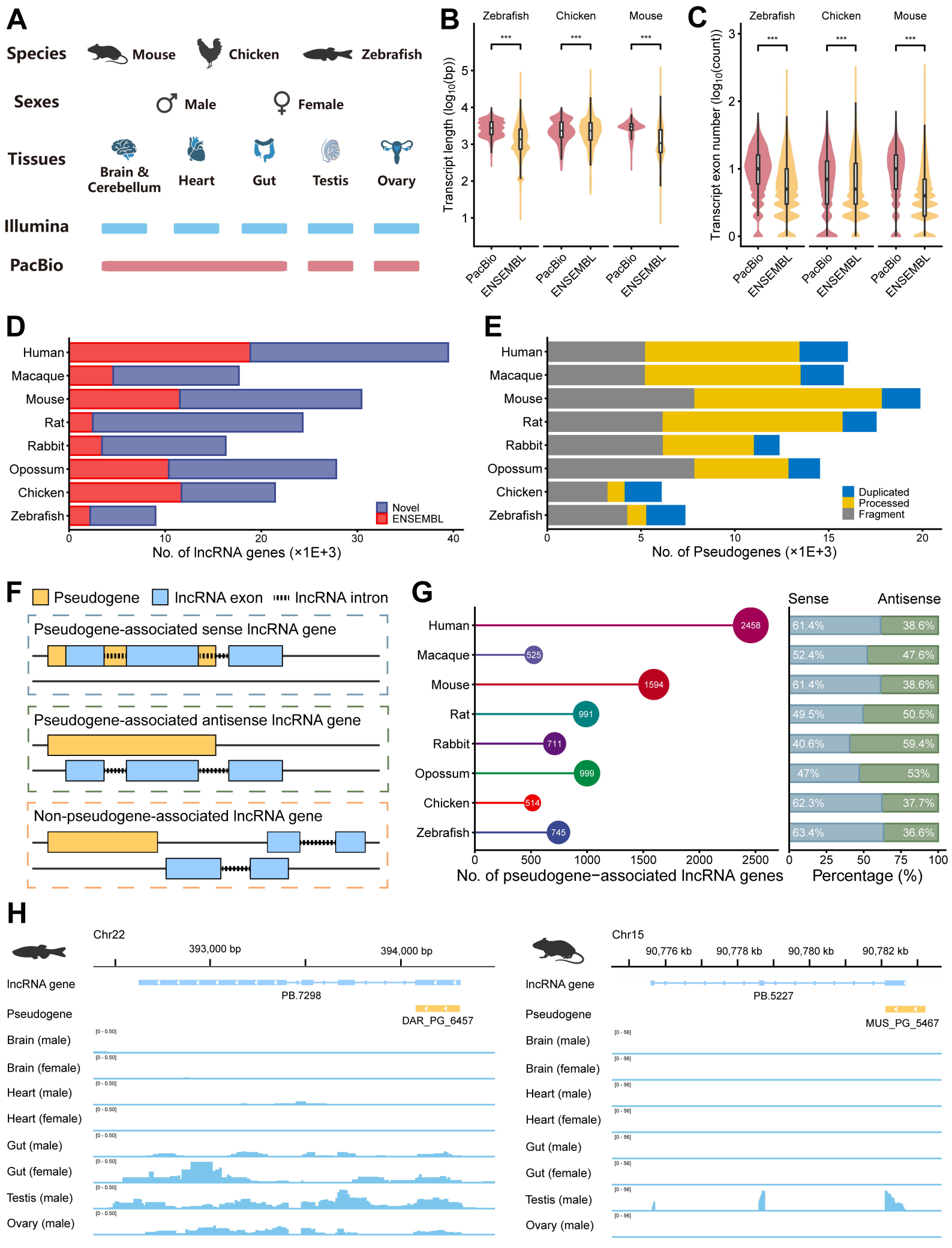
Table S3 Sample information of the public single-cell/nucleus RNA sequencing data

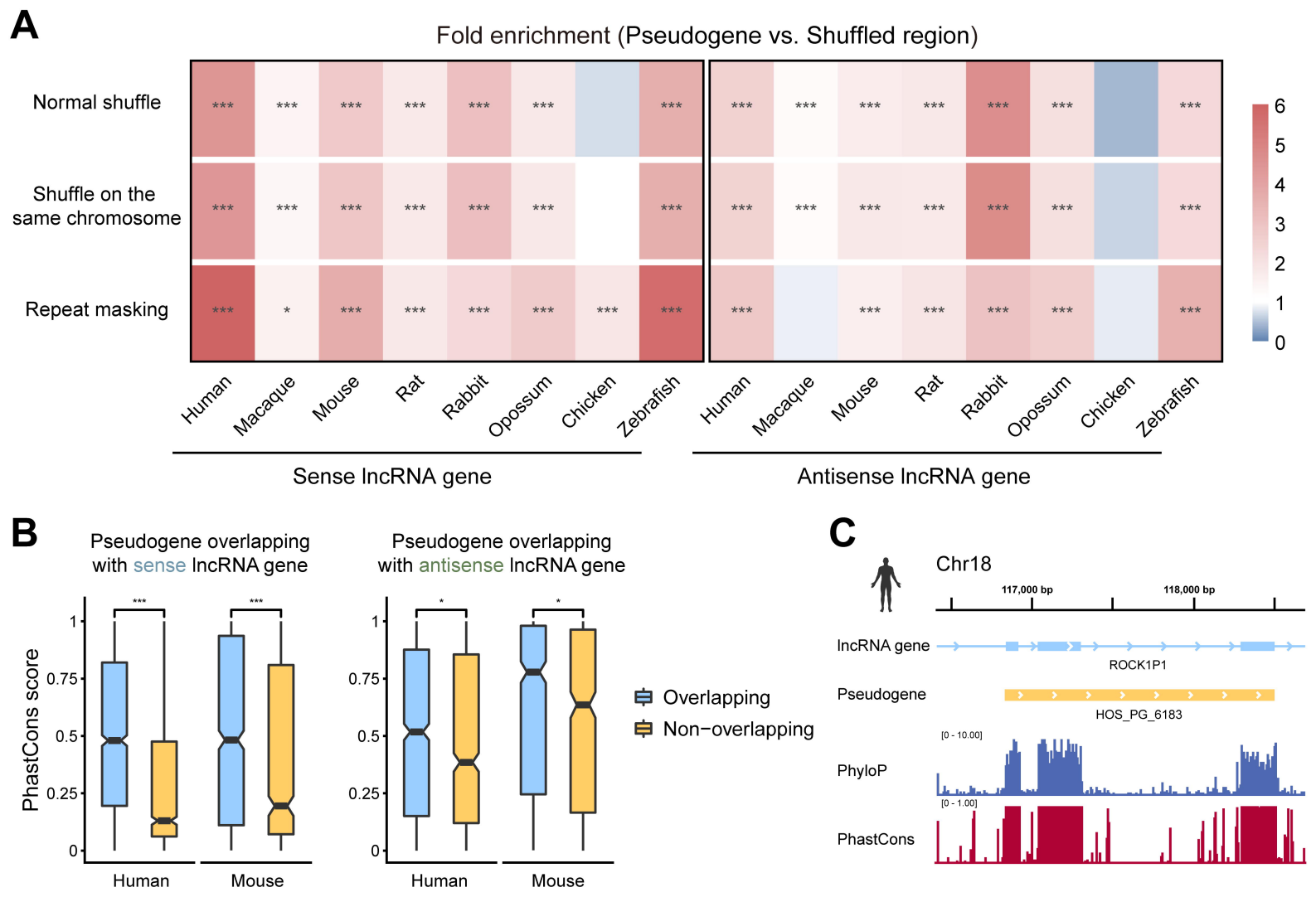
Table S4 Abbreviations, marker genes, and number for each cell type in each tissue

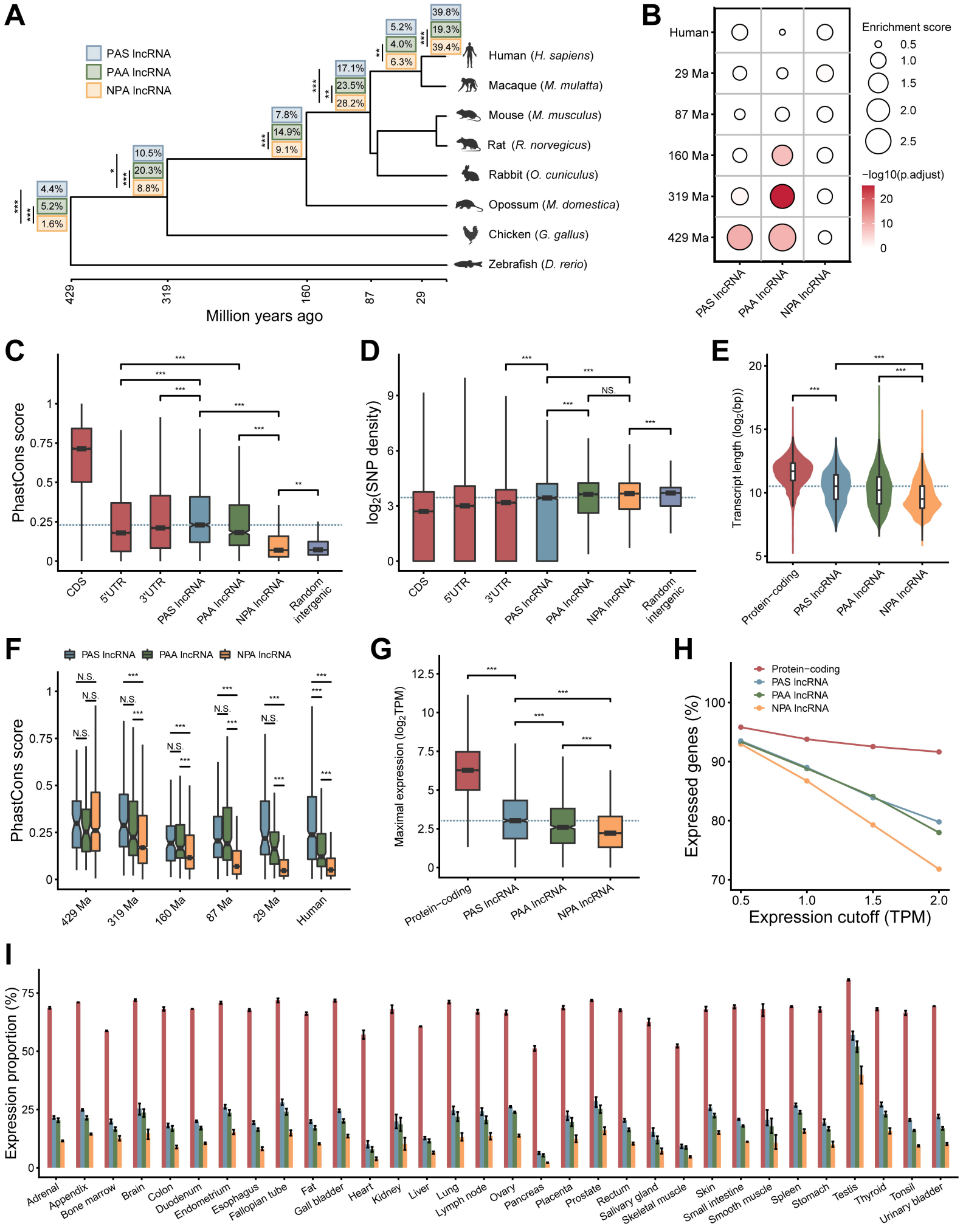
Table S5 Aging-related differential expression of pseudogene-associated lncRNA genes in human

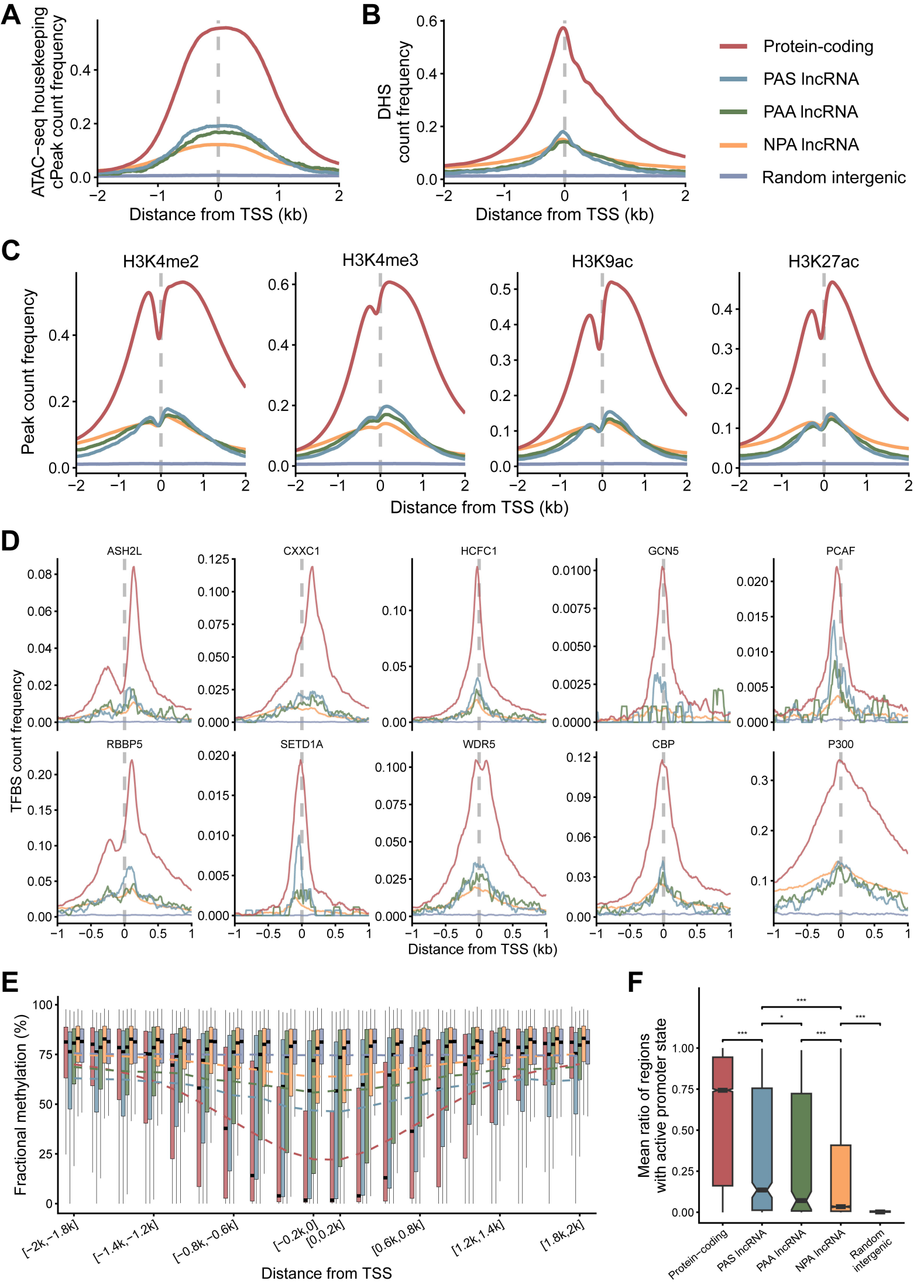
Table S6 Aging-related differential expression of pseudogene-associated lncRNA genes in mouse

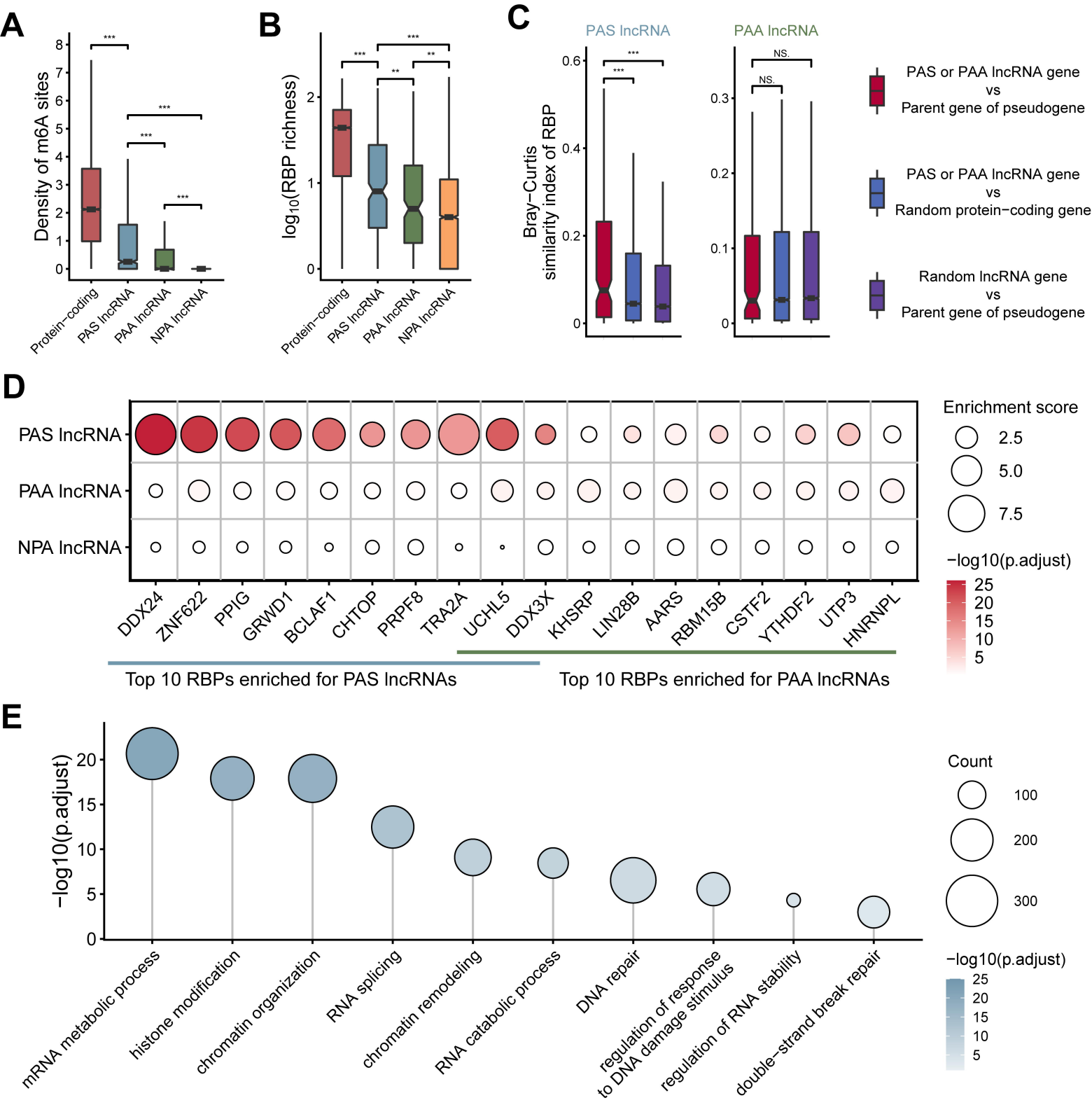
Table S7 Primers for quantitative real-time PCR

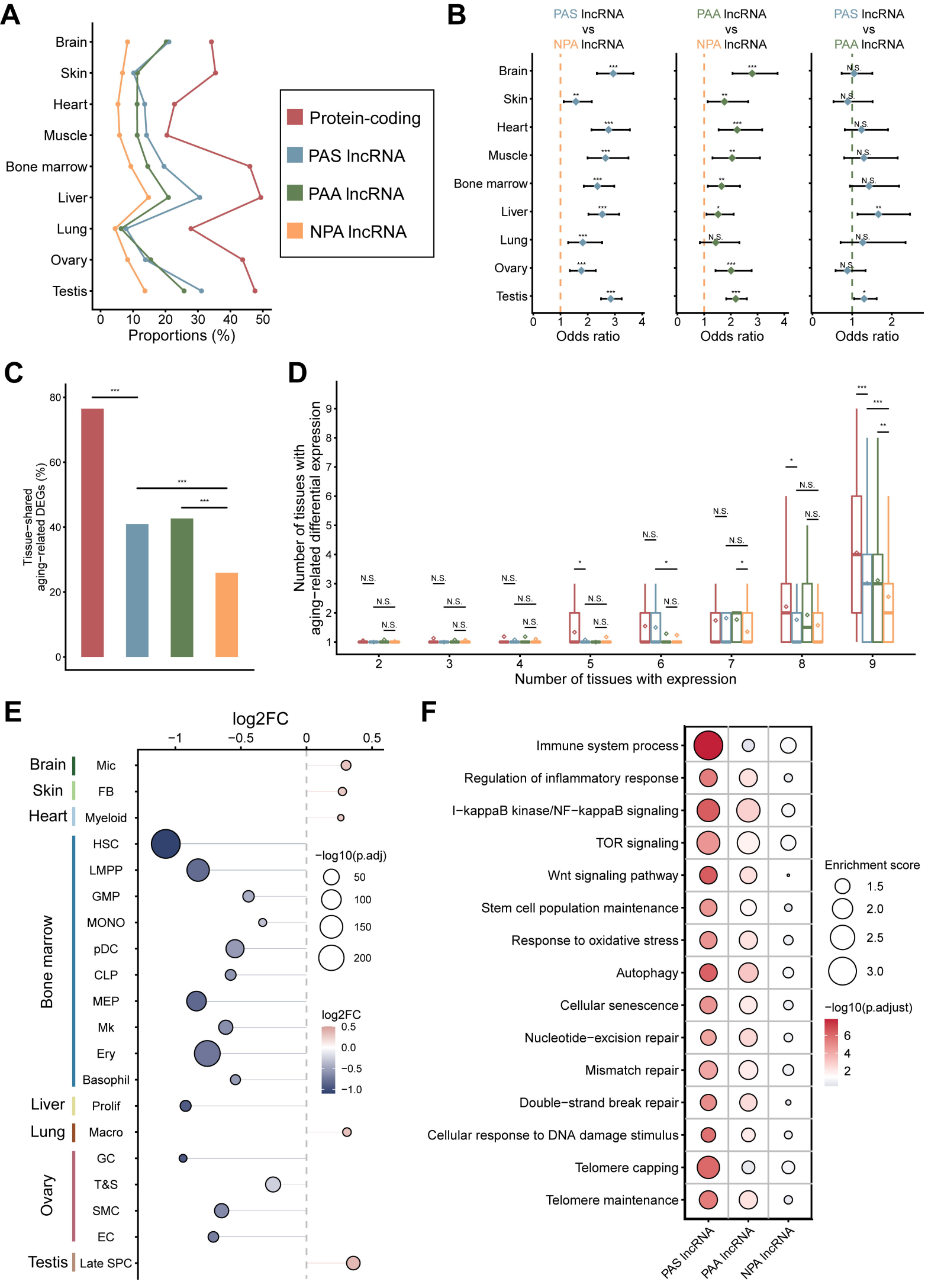

A
Species
Mouse
Chicken
Zebrafish
Sexes
Male
Female
Tissues
Brain & Cerebellum
Heart
Gut
Testis
Ovary
Illumina
PacBio
B
Zebrafish
Chicken
Mouse
Transcript length (log10(bp))
PacBio
ENSEMBL
C
Transcript exon number (log10(count))
D
Human
Macaque
Mouse
Rat
Rabbit
Opossum
Chicken
Zebrafish
Novel
ENSEMBL
No. of lncRNA genes (×1E+3)
E
Duplicated
Processed
Fragment
No. of Pseudogenes (×1E+3)
F
Pseudogene
lncRNA exon
lncRNA intron
Pseudogene-associated sense lncRNA gene
Pseudogene-associated antisense lncRNA gene
Non-pseudogene-associated lncRNA gene
G
Sense
Antisense
2458
525
1594
991
711
999
514
745
61.4%
38.6%
52.4%
47.6%
61.4%
38.6%
49.5%
50.5%
40.6%
59.4%
47%
53%
62.3%
37.7%
63.4%
36.6%
No. of pseudogene−associated lncRNA genes
Percentage (%)
H
Chr22
393,000 bp
394,000 bp
lncRNA gene
PB.7298
Pseudogene
DAR_PG_6457
Brain (male)
Brain (female)
Heart (male)
Heart (female)
Gut (male)
Gut (female)
Testis (male)
Ovary (male)
Chr15
90,776 kb
90,778 kb
90,780 kb
90,782 kb
PB.5227
MUS_PG_5467

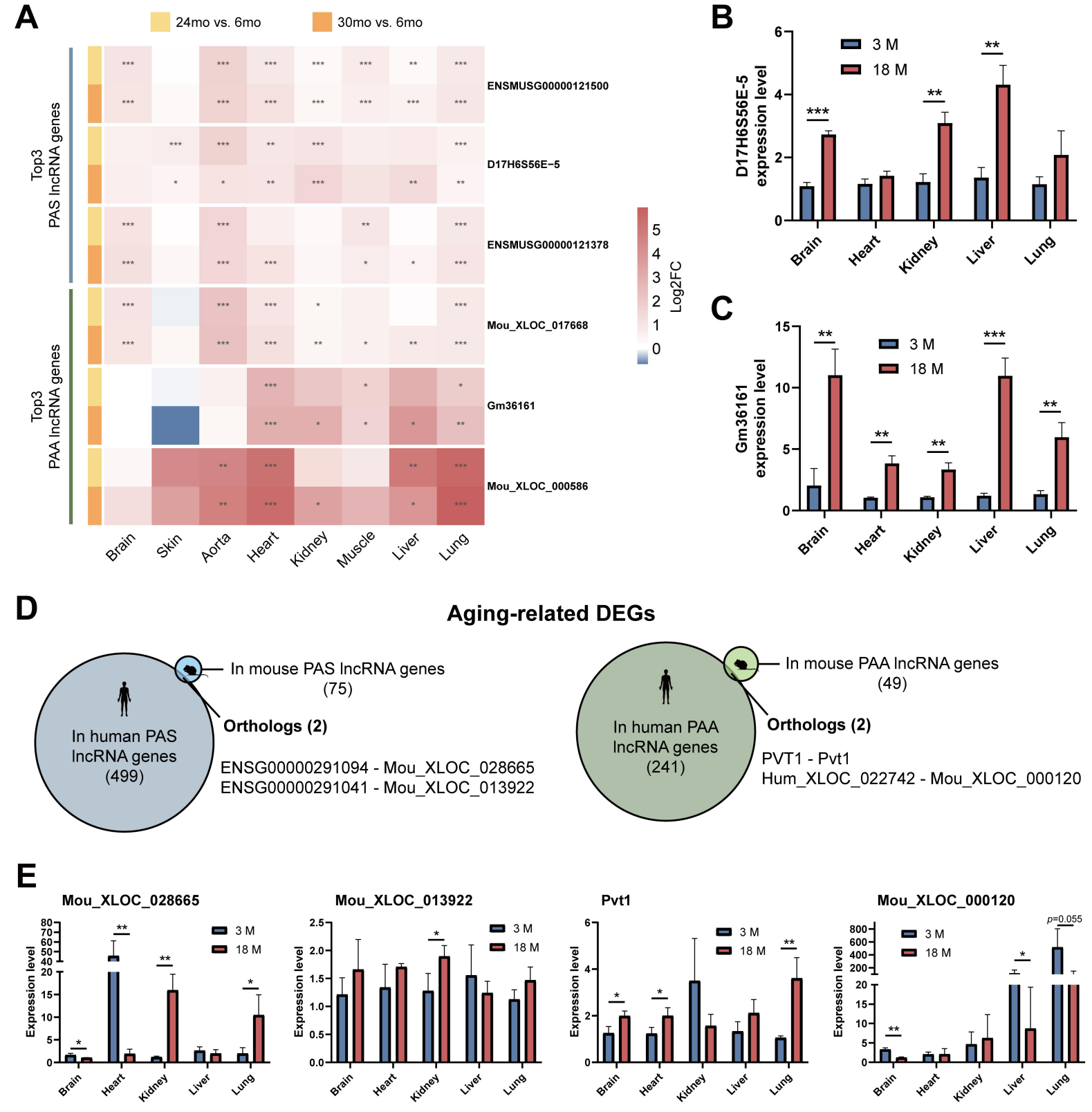

A
Fold enrichment (Pseudogene vs. Shuffled region)
Normal shuffle
Shuffle on the same chromosome
Repeat masking
Human
Macaque
Mouse
Rat
Rabbit
Opossum
Chicken
Zebrafish
Sense lncRNA gene
Antisense lncRNA gene
6
5
4
3
2
1
0
B
Pseudogene overlapping with sense lncRNA gene
Pseudogene overlapping with antisense lncRNA gene
PhastCons score
Human
Mouse
Overlapping
Non−overlapping
C
Chr18
117,000 bp
118,000 bp
lncRNA gene
ROCK1P1
Pseudogene
HOS_PG_6183
PhyloP
PhastCons


A
PAS lncRNA
PAA lncRNA
NPA lncRNA
Human (H. sapiens)
Macaque (M. mulatta)
Mouse (M. musculus)
Rat (R. norvegicus)
Rabbit (O. cuniculus)
Opossum (M. domestica)
Chicken (G. gallus)
Zebrafish (D. rerio)
Million years ago
B
Enrichment score
−log10(p.adjust)
Human
29 Ma
87 Ma
160 Ma
319 Ma
429 Ma
PAS lncRNA
PAA lncRNA
NPA lncRNA
C
PhastCons score
CDS
5'UTR
3'UTR
PAS lncRNA
PAA lncRNA
NPA lncRNA
Random intergenic
D
log2(SNP density)
E
Transcript length (log2(bp))
Protein-coding
F
429 Ma
319 Ma
160 Ma
87 Ma
29 Ma
Human
G
Maximal expression (log2TPM)
H
Expressed genes (%)
Expression cutoff (TPM)
I
Expression proportion (%)
Adrenal
Appendix
Bone marrow
Brain
Colon
Duodenum
Endometrium
Esophagus
Fallopian tube
Fat
Gall bladder
Heart
Kidney
Liver
Lung
Lymph node
Ovary
Pancreas
Placenta
Prostate
Rectum
Salivary gland
Skeletal muscle
Skin
Small intestine
Smooth muscle
Spleen
Stomach
Testis
Thyroid
Tonsil
Urinary bladder


A
ATAC−seq housekeeping cPeak count frequency
Distance from TSS (kb)
B
DHS count frequency
Distance from TSS (kb)
Protein-coding
PAS lncRNA
PAA lncRNA
NPA lncRNA
Random intergenic
C
H3K4me2
H3K4me3
H3K9ac
H3K27ac
Peak count frequency
Distance from TSS (kb)
D
ASH2L
CXXC1
HCFC1
GCN5
PCAF
RBBP5
SETD1A
WDR5
CBP
P300
TFBS count frequency
Distance from TSS (kb)
E
Fractional methylation (%)
[−2k,−1.8k]
[−1.4k,−1.2k]
[−0.8k,−0.6k]
[−0.2k,0]
[0,0.2k]
[0.6k,0.8k]
[1.2k,1.4k]
[1.8k,2k]
Distance from TSS
F
Mean ratio of regions with active promoter state
Protein−coding
PAS lncRNA
PAA lncRNA
NPA lncRNA
Random intergenic


A
Density of m6A sites
Protein−coding
PAS lncRNA
PAA lncRNA
NPA lncRNA
B
log10(RBP richness)
C
PAS lncRNA
PAA lncRNA
Bray−Curtis similarity index of RBP
NS.
PAS or PAA lncRNA gene vs Parent gene of pseudogene
PAS or PAA lncRNA gene vs Random protein−coding gene
Random lncRNA gene vs Parent gene of pseudogene
D
PAS lncRNA
PAA lncRNA
NPA lncRNA
DDX24
ZNF622
PPIG
GRWD1
BCLAF1
CHTOP
PRPF8
TRA2A
UCHL5
DDX3X
KHSRP
LIN28B
AARS
RBM15B
CSTF2
YTHDF2
UTP3
HNRNPL
Top 10 RBPs enriched for PAS lncRNAs
Top 10 RBPs enriched for PAA lncRNAs
Enrichment score
−log10(p.adjust)
E
−log10(p.adjust)
mRNA metabolic process
histone modification
chromatin organization
RNA splicing
chromatin remodeling
RNA catabolic process
DNA repair
regulation of response to DNA damage stimulus
regulation of RNA stability
double−strand break repair
Count


A
Brain
Skin
Heart
Muscle
Bone marrow
Liver
Lung
Ovary
Testis
Proportions (%)
Protein-coding
PAS lncRNA
PAA lncRNA
NPA lncRNA
B
PAS lncRNA vs NPA lncRNA
PAA lncRNA vs NPA lncRNA
PAS lncRNA vs PAA lncRNA
Odds ratio
N.S.
C
Tissue-shared aging-related DEGs (%)
D
Number of tissues with aging-related differential expression
Number of tissues with expression
E
log2FC
−log10(p.adj)
Mic
FB
Myeloid
HSC
LMPP
GMP
MONO
pDC
CLP
MEP
Mk
Ery
Basophil
Prolif
Macro
GC
T&S
SMC
EC
Late SPC
F
Immune system process
Regulation of inflammatory response
I-kappaB kinase/NF-kappaB signaling
TOR signaling
Wnt signaling pathway
Stem cell population maintenance
Response to oxidative stress
Autophagy
Cellular senescence
Nucleotide-excision repair
Mismatch repair
Double-strand break repair
Cellular response to DNA damage stimulus
Telomere capping
Telomere maintenance
PAS lncRNA
PAA lncRNA
NPA lncRNA
Enrichment score
−log10(p.adjust)


A
24mo vs. 6mo
30mo vs. 6mo
Top3 PAS lncRNA genes
Top3 PAA lncRNA genes
ENSMUSG00000121500
D17H6S56E-5
ENSMUSG00000121378
Mou_XLOC_017668
Gm36161
Mou_XLOC_000586
Brain
Skin
Aorta
Heart
Kidney
Muscle
Liver
Lung
Log2FC
B
3 M
18 M
D17H6S56E-5 expression level
C
Gm36161 expression level
D
Aging-related DEGs
In mouse PAS lncRNA genes (75)
In human PAS lncRNA genes (499)
Orthologs (2)
ENSG00000291094 - Mou_XLOC_028665
ENSG00000291041 - Mou_XLOC_013922
In mouse PAA lncRNA genes (49)
In human PAA lncRNA genes (241)
Orthologs (2)
PVT1 - Pvt1
Hum_XLOC_022742 - Mou_XLOC_000120
E
Mou_XLOC_028665
Mou_XLOC_013922
Pvt1
Mou_XLOC_000120
Expression level
p=0.055