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], Hao Zuo[1], Lu Chen[1], Meng-Wei Shi[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).

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

88 modifications, thus playing an important role in stem cell differentiation [16]. In
89 addition, lncRNAs transcribed from the *AGPG* pseudogene have been reported to
90 promote glycolysis and cell cycle progression in human [17]. The lncRNAs produced
91 by the pseudogenes *PCNAP1* and *PTENP1* play a role in the development of
92 hepatocellular carcinoma and aortic aneurysm by regulating the expression levels of
93 the parent genes *PCNA* and *PTEN*, respectively [18,19]. Together, pseudogenes can
94 be transcribed into lncRNAs, which are involved in various regulatory functions.
95 Accordingly, pseudogenes, which are often regarded as non-functional remnants of
96 coding genes, may represent an appropriate substrate for the origin of functional
97 lncRNA genes.

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

110

111 ## Results

112 ### Identification of pseudogene-associated lncRNA genes across eight vertebrate
113 genomes

114 Long-read transcriptome sequencing has proven to be accurate and effective in
115 improving genome annotation and discovering new genes [20,21]. To accurately

116 characterize lncRNA genes, we performed PacBio isoform sequencing (Iso-seq), as
117 well as RNA-seq to correct for Iso-seq data, for the adult female and male tissues
118 (brain, heart, gut, and gonad) of three representative vertebrate models, including
119 zebrafish (teleost), chicken (bird), and mouse (mammal) (**Figure 1**A). As expected,
120 the transcripts detected by PacBio Iso-seq were longer in length and had a higher
121 number of exons than the Ensembl annotated transcripts (Figure 1B, C), suggesting
122 that our home-made datasets produced high-quality genome annotations. To broaden
123 the scope for comparative studies, we also included five additional mammals (human,
124 macaque, rat, rabbit, and opossum), and a total of eight species were covered for
125 subsequent analysis. To ensure comprehensive lncRNA annotation, we integrated the
126 Ensembl lncRNA annotation with novel lncRNAs from our home-made Iso-seq
127 datasets and published lncRNA sets from previous studies (Methods; Figure S1),
128 resulting in a robust annotation of 9002 (in zebrafish) to 39,461 (in human) lncRNA
129 genes for each species (Figure 1D). Furthermore, novel lncRNA genes annotated by
130 long-read transcriptome sequencing frequently exhibit higher transcript length, higher
131 sequence conservation, and higher expression level in comparison with those
132 annotated by short-read sequencing (Figure S2–4). These results suggest that
133 long-read transcriptome sequencing can facilitate the accurate identification of novel
134 lncRNA genes with more functional properties.

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

145 processed pseudogenes (Figure S5). Additionally, the number and density of
146 processed pseudogenes in chicken and zebrafish are lower than in mammals (Figure
147 S6). Chicken CR1 reverse transcriptase is unlikely to copy polyadenylated mRNAs,
148 whereas polyadenylated tails are potential substrates for mammalian L1 reverse
149 transcriptase [25,26]. This may explain the differences in the abundance of processed
150 pseudogenes between chicken and mammals. In zebrafish, the genomic coverage of
151 retrotransposon elements is only a quarter of that in human, which could account for
152 the lack of processed pseudogenes [27].

153 Next, we identified the pseudogene-associated lncRNA genes based on two
154 different overlap patterns (Methods; Figure 1F, G; Table S1), and found 514 (in
155 chicken) to 2458 (in human) lncRNA genes, accounting for 2.4% to 8.3% of all the
156 lncRNA genes in each species. A robust positive correlation was identified between
157 the number of pseudogene-associated lncRNA genes and the total number of lncRNA
158 genes across diverse species (Figure S7). This finding suggests that although part of
159 the variability in the number of pseudogene-associated lncRNA genes may be
160 biologically meaningful, much is likely to be explained by unequal size of the
161 lncRNA gene repertoire. Consequently, factors influencing the size of the lncRNA
162 annotation, including the quality of genome sequence and annotation, as well as the
163 amount and sequencing depth of RNA-seq data used to identify novel lncRNA genes
164 [28], affect the number of pseudogene-associated lncRNA genes. Additionally, 41.7%
165 (in human) to 75.2% (in rat) of the pseudogene-associated lncRNA genes in each
166 species were novel lncRNA genes (Figure S8). This highlights the significance of
167 comprehensive lncRNA annotation. For example, zebrafish novel lncRNA gene
168 *PB.7298* and mouse novel lncRNA gene *PB.5227* detected by Iso-seq overlapped with
169 the pseudogene and were identified as pseudogene-associated lncRNA genes, with the
170 first exon of both derived from the pseudogene (Figure 1H). *PB.7298* was found to be
171 expressed in the gut and gonads, whereas *PB.5227* was only expressed in the testis.

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

173 To understand the contribution of pseudogenes to the makeup and function of lncRNA

174 genes, we evaluated the enrichment of lncRNA genes at pseudogenes (Methods) and
175 found that both sense and antisense lncRNA genes were enriched at pseudogenes
176 compared to shuffled regions in mammals and zebrafish (empirical $P \leq 0.001$)
177 (**Figure 2**A). After excluding the potential effects of chromosomal biases or repetitive
178 sequences (Methods), similar results were still observed (Figure 2A). We further
179 evaluated the enrichment of lncRNA genes at each class of pseudogenes, and found
180 that both duplicated and processed pseudogenes showed high enrichment of lncRNA
181 genes in comparison to shuffled regions (Figure S9). These findings imply that
182 pseudogenes act as relics of protein-coding gene copies to provide favorable
183 substrates for the origin of functional lncRNA genes, which are retained by purifying
184 selection during evolution.

185 To further confirm whether the origin of lncRNA genes is functionally gained or
186 lost for pseudogene regions. We evaluated the sequence conservation of overlapping
187 and non-overlapping regions within the pseudogenes overlapping with sense and
188 antisense lncRNA genes using PhastCons score in human and mouse. We found that
189 regions within pseudogenes that overlap with exons of lncRNA genes are more
190 conserved than non-overlapping portions (two-sided Wilcoxon rank sum test,
191 $*P < 0.05$, $***P < 0.001$) (Figure 2B), suggesting that these pseudogene regions
192 acquire enhanced functional properties through the origin of sense or antisense
193 lncRNA genes. For example, within the pseudogene HOS_PG_6183, which overlaps
194 with the human lncRNA gene *ROCK1P1*, both PhastCons and PhyloP scores showed
195 higher sequence conservation in the regions overlapping the *ROCK1P1* exons (Figure
196 2C).

197 **Enhanced functional attributes of pseudogene-associated lncRNA genes in**
198 **sequence**

199 The enrichment of lncRNA genes at pseudogenes implies that pseudogenes may have
200 endowed pseudogene-associated lncRNA genes with functionality. To explore the
201 functional potential of pseudogene-associated lncRNA genes, we conducted
202 comparative analyses with non-pseudogene-associated (NPA) counterparts, focusing

203 on evolutionary age, sequence conservation, intrapopulation variation, and structural
204 attributes.

205 We estimated the evolutionary age of human and mouse lncRNA genes using 1:1
206 orthologous lncRNA genes based on the phylogenetic relationships of each species
207 (Methods; Table S2). Although large proportions of lncRNA genes were
208 species-specific in human (38.9%) and mouse (35.6%), both pseudogene-associated
209 sense (PAS) and pseudogene-associated antisense (PAA) lncRNA genes had higher
210 proportions of genes in older age groups, and lower proportions of genes in younger
211 age groups than NPA lncRNA genes (**Figure 3**A; Figure S10). Next, we analyzed the
212 enrichment of age groups for each category of lncRNA genes using all lncRNA genes
213 as background genes. Take human for example, we found that NPA lncRNA genes
214 were not enriched in any of the evolutionary age groups (Figure 3B). In contrast, PAS
215 lncRNA genes were enriched in 429 million years ago (Ma), and PAA lncRNA genes
216 were enriched in 429 Ma, 319 Ma, and 160 Ma. A similar phenomenon was observed
217 in mouse (Figure S11). Taken together, our analysis suggests that
218 pseudogene-associated lncRNA genes exhibit higher evolutionary conservation,
219 which implies their greater functionality.

220 To further reveal the functional potential of pseudogene-associated lncRNA genes,
221 we evaluated the cross-species DNA sequence conservation of these lncRNA genes
222 (Figure 3C; Figure S12; Table S2). We found that the exon sequence conservation of
223 both PAS (median ~0.23 in human and ~0.24 in mouse) and PAA lncRNA genes
224 (median ~0.18 in human and ~0.29 in mouse) was lower than that of the coding
225 sequence (CDS) of protein-coding genes, but higher than that of NPA lncRNA genes
226 (median ~0.07 in human and ~0.11 in mouse) and random intergenic regions (median
227 ~0.07 in human and ~0.09 in mouse), and even higher than that of the 5' untranslated
228 regions (UTRs) of protein-coding genes in both human and mouse (two-sided
229 Wilcoxon rank sum test, $P < 0.001$). Conservation beyond the non-coding regions of
230 the exons of protein-coding genes, demonstrating greater functionality of
231 pseudogene-associated lncRNA genes. To ensure that the higher conservation of

232 pseudogene-associated lncRNA genes compared to NPA counterparts was not due to
233 their older age, we evaluated the conservation of lncRNA genes in each age group
234 (Figure 3F; Figure S13). We found that each category of lncRNA genes is highly
235 conserved in the oldest age group, but both PAS and PAA lncRNA genes remain
236 consistently more conserved than NPA counterparts in other age groups, suggesting
237 stronger evolutionary constraints on pseudogene-associated lncRNA genes at each
238 age.

239 Next, we investigated the intrapopulation variation of pseudogene-associated
240 lncRNA genes in human [29]. Although the derived allele frequencies of single
241 nucleotide polymorphisms (SNP) found in the exons of pseudogene-associated
242 lncRNA genes were not different from those found in the exons of NPA lncRNA
243 genes (two-sided unpaired T test, $P > 0.05$) (Figure S14), the SNP density of exons of
244 PAS lncRNA genes was lower than that of exons of NPA lncRNA genes as well as
245 random intergenic regions (two-sided Wilcoxon rank sum test, $P < 0.001$) (Figure 3D).
246 This suggests that PAS lncRNA genes may be subject to slightly stronger
247 intrapopulation purifying selection than NPA lncRNA genes. In addition, the
248 Combined Annotation-Dependent Depletion (CADD) score [30] is a deleteriousness
249 score for human variants. We found that CADD scores for SNPs in exons of both PAS
250 and PAA lncRNA genes were higher than those in exons of NPA lncRNA genes
251 (two-sided Wilcoxon rank sum test, $P < 0.001$) (Figure S15), suggesting that variants
252 in pseudogene-associated lncRNA genes are more pathogenic than those in NPA
253 lncRNA genes. These results again implied the potential functionality of
254 pseudogene-associated lncRNA genes.

255 Since the functionality of genes is often related to the structural attributes of their
256 transcripts [1,31], we examined the structural attributes of the transcripts of
257 pseudogene-associated lncRNA genes (Table S2). We found that both PAS and PAA
258 lncRNA genes had higher transcript lengths and exon numbers than NPA lncRNA
259 genes (two-sided Wilcoxon rank sum test, $P < 0.001$; two-sided unpaired T test, $P <$
260 0.001) (Figure 3E; Figure S16, 17), suggesting that the transcripts of

261 pseudogene-associated lncRNA genes may carry more RNA structural domains,
262 thereby favoring their interactions with other nucleic acid or protein molecules [32].
263 Altogether, these findings in sequence revealed that pseudogene-associated
264 lncRNA genes possess enhanced functional attributes compared to NPA counterparts.

265 **Heightened functional attributes of pseudogene-associated lncRNA genes in**
266 **expression**

267 To further explore the functionality of pseudogene-associated lncRNA genes in
268 expression, we conducted comparisons of their expression profiles and epigenetic
269 features, including chromatin accessibility, histone modifications, and DNA
270 methylation, near transcription start sites (TSSs) with protein-coding genes and NPA
271 lncRNA genes.
272 Utilizing RNA-seq data from 32 human tissues, we determined expression profiles
273 across tissues and found that the expression level of both PAS and PAA lncRNA genes,
274 compared with NPA lncRNA genes, was closer to protein-coding genes (two-sided
275 Wilcoxon rank sum test, $P < 0.001$) (Figure 3G; Figure S18; Table S2). Moreover, the
276 expression ratios of both PAS and PAA lncRNA genes were higher than that of NPA
277 lncRNA genes, and closer to that of protein-coding genes at all expression cutoffs
278 (Figure 3H) and in all 32 tissues (Figure 3I). Notably, PAS lncRNA genes show the
279 highest expression activity. We also observed a conserved high expression activity of
280 PAS lncRNA genes in mouse (Figure S19–22). In addition, each category of lncRNA
281 genes both showed tissue-specific expression patterns (Figure S23; Table S2),
282 supporting the high tissue-specificity of lncRNA gene expression [33,34]. Meanwhile,
283 both PAS and PAA lncRNA genes (median: 7 and 6 tissues), as well as coding genes,
284 exhibited broader tissue expression profiles than NPA lncRNA genes (median: 4
285 tissues) in human (two-sided Wilcoxon rank sum test, $P < 0.001$) (Figure S23),
286 demonstrating their potential functional versatility across tissues.
287 Chromatin accessibility plays a crucial role in regulating gene expression by
288 influencing the accessibility of DNA to transcriptional machinery and regulatory
289 proteins. Therefore, to elucidate the causes of active expression of

290 pseudogene-associated lncRNA genes at the epigenomic level, we characterized
291 chromatin accessibility near the transcription start sites (TSSs) in human. Compared
292 to NPA lncRNA genes, both PAS and PAA lncRNA genes showed more housekeeping
293 chromatin open regions by ATAC-seq, and PAS lncRNA genes also showed higher
294 DNase accessibility by DNase-seq (**Figure 4**A, B), demonstating higher chromatin
295 accessibility at proximal promoters of pseudogene-associated lncRNA genes.

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

314 Besides histone modifications, DNA methylation are also important regulators of
315 gene expression. We thus examined DNA methylation in each 200 bp window within
316 a total of 4kb regions upstream and downstream of the gene TSSs, and found that the
317 DNA methylation of both PAS and PAA lncRNA genes was lower than that of NPA
318 lncRNA genes in each window (two-sided Wilcoxon rank sum test, $P < 0.001$) (Figure

319 4E), and reached the lowest level at TSS, indicating that the pseudogene-associated
320 lncRNA genes could more effectively avoid the silencing of gene expression caused
321 by DNA methylation in the promoter region.

322 As an integration of chromatin accessibility, histone modifications, DNA
323 methylation, and other epigenetic feature, chromHMM chromatin states serve as a
324 molecular framework influencing gene expression patterns [38]. We found that the
325 ratio of active promoter state regions in the proximal promoters of both PAS and PAA
326 lncRNA genes was higher than that in the proximal promoters of NPA lncRNA genes
327 as well as in the random intergenic region (two-sided Wilcoxon rank sum test, $P <$
328 0.001) (Figure 4F). These results are consistent with their epigenomic
329 characterization.

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

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

336 lncRNAs usually function as regulators of gene expression [39]. In general, lncRNAs
337 often exert their regulatory roles by interacting with RNA-binding proteins (RBPs)
338 [40,41]. They regulate gene expression at the transcriptional level by recruiting
339 transcription factors and chromatin modifiers, or at the post-transcriptional level by
340 competing with mRNAs for binding to RNA stabilizing/destabilizing factors [2,42,43].
341 Therefore, to explore the potential function of pseudogene-associated lncRNA genes,
342 we scrutinized their features associated with RNA-protein binding in human.

343 N6-methyladenosine (m6A) is a prevalent and significant RNA modification
344 known to influence the binding between RNAs and proteins [44,45]. We investigated
345 the m6A modification in pseudogene-associated lncRNAs and found that the density
346 of m6A sites in both PAS and PAA lncRNAs was higher than that in NPA lncRNAs
347 (two-sided Wilcoxon rank sum test, $P < 0.001$) (**Figure 5**A; Table S2), suggesting a

348 more complex binding pattern between pseudogene-associated lncRNAs and proteins.
349 Next, we directly characterized the binding of pseudogene-associated lncRNAs to
350 RBPs (Table S2). In transcripts with RBP binding sites, the richness of RBPs bound to
351 both PAS and PAA lncRNAs was higher than those bound to NPA lncRNAs
352 (two-sided Wilcoxon rank sum test, $^{**}P < 0.01$, $^{***}P < 0.001$) (Figure 5B).
353 Furthermore, the diversity of RBPs bound to PAS lncRNAs was also higher than
354 those bound to NPA lncRNAs (Figure S24). To exclude the effect of gene expression
355 levels on the detection of m6A sites as well as RBP binding sites, we evaluated these
356 features associated with RNA-protein binding in four expression level groups and
357 found similar results (Figure S25–27). These results suggest that
358 pseudogene-associated lncRNAs bind more abundant RBPs than NPA lncRNAs,
359 indicating that pseudogene-associated lncRNAs are greater regulators of gene
360 expression compared to NPA counterparts.

361 Although both PAS and PAA lncRNAs showed advantages in protein-binding
362 capacity, PAS lncRNAs exhibited higher m6A site density as well as more abundant
363 and diverse RBP binding sites compared to PAA lncRNAs (Figure 5A, B; Figure
364 S24). Since PAS lncRNAs have homologous sequences of transcripts from parent
365 genes of pseudogenes, whereas PAA lncRNAs have potential complementary
366 sequences of transcripts from parent genes of pseudogenes, we hypothesized that the
367 abundant RBP binding sites of PAS lncRNAs are conferred by parent genes of
368 pseudogenes. To test this hypothesis, we calculated the Bray-Curtis similarity of RBP
369 binding sites between each category of pseudogene-associated lncRNAs and
370 transcripts from parent genes of pseudogenes. A higher RBP similarity was identified
371 between PAS lncRNAs and transcripts from parent genes of pseudogenes compared to
372 gene pairs used as controls (two-sided Wilcoxon rank sum test, $P < 0.001$) (Figure
373 5C). However, the RBP similarity between PAA lncRNAs and transcripts from parent
374 genes of pseudogenes did not differ from that observed in gene pairs used as controls
375 (Figure 5C). These findings suggest that only PAS lncRNAs inherit the RBP binding
376 sites from parent genes of pseudogenes, thereby demonstrating greater proficiency in

377 protein-binding.

378 Furthermore, we analyzed the enrichment of RBPs for each category of lncRNAs
379 using all lncRNAs as background. We found that NPA lncRNAs were not enriched for
380 RBPs, whereas PAS lncRNAs and PAA lncRNAs were enriched for 104 and 12 RBPs,
381 respectively (Figure S28). These RBPs include a number of gene expression
382 regulators (Figure 5D). The top 10 RBPs enriched by PAS lncRNAs include
383 transcriptional regulators (ZNF662, GRWD1, BCLAF1, CHTOP, DDX3X), RNA
384 splicing regulators (PPIG, PRPF8, TRA2A), deubiquitinase (UCHL5), and mRNA
385 stabilizing factor (DDX24). Meanwhile, PAA lncRNAs were also enriched for
386 DDX3X and UCHL5, as well as other gene expression regulators such as regulators
387 of RNA splicing and processing (RBM15B, CSTF2, HNRNPL), translation regulator
388 (LIN28B), and mRNA degradation factors (KHSRP, YTHDF2). These results suggest
389 that pseudogene-associated lncRNAs may be involved in the regulation of gene
390 expression at multiple levels by binding to RBPs. Moreover, DDX24 [46] and
391 DDX3X [47,48] were found to be closely associated with DNA repair, indicating a
392 potential involvement of pseudogene-associated lncRNAs in DNA repair.

393 Co-expression can indicate functional relevance or regulatory relationships
394 between lncRNAs and protein-coding genes [28]. To infer the putative functions of
395 pseudogene-associated lncRNA genes, we identified co-expressed protein-coding
396 genes for each lncRNA gene in human and mouse (Methods). We further analyzed the
397 enrichment of co-expressed genes to identify co-expressed genes more closely
398 associated with the function of each category of lncRNA genes. We found that both
399 PAS and PAA lncRNA genes were enriched for more co-expressed protein-coding
400 genes compared to NPA lncRNA genes in both human and mouse (Figure S29, 30).
401 Next, co-expressed genes enriched by each category of lncRNA genes were analyzed
402 for Gene Ontology (GO) enrichment analysis. The co-expressed genes enriched by
403 NPA lncRNA genes were not enriched in any of the entries in human and were
404 enriched in mitosis-related entries in mouse (Figure S31). Whereas the co-expressed
405 genes enriched by PAS lncRNA genes were enriched in functions related to the

406 regulation of gene expression, such as histone modification, chromatin organization,
407 chromatin remodeling, mRNA metabolic process, RNA catabolic process, and
408 regulation of RNA stability in both human and mouse (Figure 5E; Figure S32).
409 Meanwhile, the co-expressed genes enriched by PAA lncRNA genes were enriched in
410 similar functions related to gene expression regulation in mouse (Figure S33, 34).
411 These results again implies the important role of pseudogene-associated lncRNA
412 genes in gene expression regulation. Enrichment analysis also revealed that
413 pseudogene-associated lncRNA genes have conserved functions in DNA repair,
414 double-strand break repair, and regulation of response to DNA damage stimulus
415 (Figure 5E; Figure S32, 34).

416 Taken together, our analyses about RBPs and co-expression collectively suggested
417 that pseudogene-associated lncRNA genes have unique advantages and conserved
418 functions in gene expression regulation and DNA repair, which may explain the high
419 conservation of pseudogene-associated lncRNA genes.

420 **Deciphering aging-related expression changes of pseudogene-associated lncRNA**
421 **genes across tissues**

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

429 After cell filtration, we retained a total of 850,860 high-quality single cells/nuclei
430 and identified 72 cell types based on canonical cell markers in all tissues (Figure S35;
431 Table S4). We found that both PAS and PAA lncRNA genes were more actively
432 expressed than NPA lncRNA genes in each cell type of each tissue (two-sided Fisher's
433 exact test, $P < 0.001$) (Figure S36, 37). Meanwhile, both PAS and PAA lncRNA genes,
434 along with coding genes, displayed more extensive cell type expression profiles than

435 NPA lncRNA genes across all tissues (two-sided Wilcoxon rank sum test, $P < 0.001$)
436 (Figure S38), indicating their potential functional versatility across cell types.
437 Subsequently, a differential expression analysis was conducted between the young
438 and old age groups in each cell type of each tissue (Table S5). We found that the
439 proportion of aging-related differentially expressed genes (DEGs) in both PAS and
440 PAA lncRNA genes was higher than that in NPA lncRNA genes in each tissue (**Figure**
441 **6**A, B), suggesting a more prominent contribution of pseudogene-associated lncRNA
442 genes to aging-related transcriptome changes and their closer association with aging.
443 Furthermore, we also observed that the close association of pseudogene-associated
444 lncRNA genes with aging was cell type-specific (Figure S39). For instance, in the
445 brain, the proportion of up-regulated genes in PAS lncRNA genes was higher than that
446 in NPA lncRNA genes in 6 out of 8 cell types, suggesting that pseudogene-associated
447 lncRNA genes respond differently to aging in different cell types of the same tissue.
448 This underscores the necessity for a detailed examination of their aging-related
449 alterations in expression at the single-cell level.
450 Among aging-related DEGs, we observed that the proportion of tissue-shared
451 DEGs (genes differentially expressed in at least two tissues) in both PAS and PAA
452 lncRNA genes was higher than that in NPA lncRNA genes (two-sided Fisher's exact
453 test, $P < 0.001$) (Figure 6C), suggesting that pseudogene-associated lncRNA genes
454 respond to aging more broadly across tissues. Considering the broader tissue
455 expression profile of pseudogene-associated lncRNA genes (Figure S23, 40), we
456 proceeded to investigate the number of tissues in which the genes exhibited
457 aging-related differential expression in each expression breadth group. Both PAS and
458 PAA lncRNA genes were found to exhibit aging-related differential expression in a
459 greater number of tissues, particularly in the group of genes expressed in nine tissues
460 (Figure 6D), suggesting that pseudogene-associated lncRNA genes are more
461 commonly involved in the aging process. One illustrative example of
462 pseudogene-associated lncRNA genes that exhibit aging-related differential
463 expression in multiple tissues is *JPX*, which demonstrates aging-related differential

464 expression in 20 cell types across 8 tissues (Figure 6E). *JPX* has been demonstrated to
465 be directly related to aging and represents a potential therapeutic target for the
466 treatment of age-related atherosclerosis [49]. Overall, these findings indicate that
467 pseudogene-associated lncRNA genes exhibit a positive and extensive response to
468 aging.

469 To investigate whether pseudogene-associated lncRNA genes show a close
470 association with aging due to transcriptional regulation, we assessed the enrichment of
471 transcription factor (TF) binding sites in the proximal promoters of each category of
472 lncRNA genes in human. Both PAS and PAA lncRNA genes were enriched for the
473 binding sites of aging-related TFs (Figure S41), including NRF1 [50], E2F6, RB1
474 [51], PML [52], and SREBF2 [53]. Furthermore, the binding sites of longevity factors
475 (FOXO3, SIRT6) [54,55], rejuvenation factor (SOX5) [56], and aging modulators
476 (MTOR, ZBTB7A, CXXC1, SREBF1) [57,58] were also enriched in the proximal
477 promoters of PAS lncRNA genes. Conversely, NPA lncRNA genes were not enriched
478 for the binding sites of these aging-related TFs. These results suggest that
479 pseudogene-associated lncRNA genes are subject to transcriptional regulation of
480 aging-related TFs, thereby closely associating with aging.

481 To further explore the putative functions of pseudogene-associated lncRNA genes
482 in aging, we conducted GO enrichment analysis based on the guilt-by-association
483 method for aging-related DEGs in each category of lncRNA genes. The aging-related
484 DEGs in both PAS and PAA lncRNA genes were enriched for aging-related biological
485 processes (Figure 6F), including regulation of inflammatory response, NF-kappaB
486 signaling, TOR signaling, response to oxidative stress, cellular senescence, DNA
487 repair, and telomere maintenance. In contrast, the aging-related DEGs in NPA
488 lncRNA genes were not enriched for these aging-related entries. These findings
489 suggest that pseudogene-associated lncRNA genes regulate aging-related pathways in
490 aging.

491 To investigate whether the close association of pseudogene-associated lncRNA
492 genes with aging is conserved in mouse, we used RNA-seq data from 8 mouse tissues

493 to characterize the aging-related expression changes of pseudogene-associated
494 lncRNA genes (Methods; Table S6). We found that both PAS and PAA lncRNA genes
495 showed a higher proportion of aging-related DEGs and a broader aging-related
496 differential expression compared to NPA lncRNA genes across mouse tissues (Figure
497 S42). To experimentally validate the aging-related expression changes of
498 pseudogene-associated lncRNA genes, we quantified the expression levels of the top 3
499 PAS and PAA lncRNA genes differentially expressed in the most mouse tissues by
500 quantitative reverse transcription PCR (RT-qPCR). These pseudogene-associated
501 lncRNA genes showed aging-related differential expression in multiple tissues by
502 both RNA-seq and RT-qPCR analysis (**Figure 7**A, B, C; Figure S43). These results
503 confirm the conserved and close association between pseudogene-associated lncRNA
504 genes and aging. Furthermore, we identified two orthologous PAS lncRNA gene pairs
505 and two orthologous PAA lncRNA gene pairs that exhibited aging-related differential
506 expression in both human and mouse (Figure 7D; Figure S44). These lncRNA genes
507 may play a conserved role in aging. Among them, *PVT1* has been shown to be
508 involved in the regulation of aging-related diseases and skin aging [59,60]. We
509 validated aging-related differential expression of mouse orthologs in these
510 orthologous gene pairs by RT-qPCR (Figure 7E). Together, the characterization of
511 aging-related expression changes across tissues and species suggested that the
512 association between pseudogene-associated lncRNA genes and aging is tight and
513 conserved. Therefore, further rigorous studies of the pseudogene-associated lncRNA
514 genes identified in this study will facilitate the discovery of novel aging markers and
515 therapeutic targets for aging-related diseases in the future.
516

517 ## Discussion

518 lncRNA-related research is growing by the day, but there is still an intense debate
519 about the proportion of functional lncRNA genes in the total lncRNA gene collection
520 [6]. Studies of eukaryotic evolutionary dynamics and transcription machinery have
521 shown that lncRNA annotation contain a large number of non-functional lncRNA

522 genes [61]. The co-occurrence of non-functional junk and functional lncRNA genes in
523 lncRNA annotation hinders the broader exploration of functional lncRNA genes.
524 Pseudogenes are often considered as non-functional sequences and thus excluded
525 from functional genomics studies. However, increasing evidence suggests that
526 pseudogenes may be favorable substrates for the origin of functional lncRNA genes
527 [16,19,62].

528 We identified pseudogene-associated lncRNA genes using genome-wide
529 pseudogenes and a comprehensive lncRNA annotation based on the Ensembl lncRNA
530 annotation, as well as novel lncRNA genes from our home-made Iso-seq datasets and
531 published lncRNA sets from previous studies. More than 40% of
532 pseudogene-associated lncRNA genes are novel lncRNA genes in each species,
533 suggesting that the comprehensiveness of the lncRNA gene repertoire is important.
534 Our human lncRNA gene annotation is not as comprehensive as the LncBook
535 database, which provides a comprehensive and high-quality collection of human
536 lncRNA genes [63] (Figure S45). However, while ensuring the systematic of
537 construction of lncRNA gene annotation and the comparability of lncRNA genes
538 across different species, the comprehensiveness of lncRNA annotation in our study is
539 comparable to that in recent studies [31,41]. The enhancement of lncRNA gene
540 annotation across species in the future will facilitate the identification of
541 pseudogene-associated lncRNA genes.

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

550 Multidimensional evidence supports the enhanced functionality of

pseudogene-associated lncRNA genes, including higher evolutionary conservation, higher exon sequence conservation, longer transcripts, more active expression, more active transcriptional regulation, and greater protein-binding capabilities compared to NPA counterparts. In order to provide a more accurate reflection of the influence of pseudogenes on lncRNA genes, a filtering process was conducted on pseudogene-associated lncRNA genes based on stricter overlap length thresholds between pseudogenes and lncRNA genes. We found that pseudogene-associated lncRNA genes showed further heightened functional potential at stricter overlap length cutoffs (Figure S46–49). Taken human PAS lncRNA genes for example, they exhibit invariant evolutionary age distribution and RBP-binding capabilities as well as higher conservation, transcript length, expression levels, and m6A site density at stricter overlap length thresholds (Figure S46). Additionally, transcriptional activity might enhance the identification criteria of pseudogene-associated lncRNA genes. However, the functional potential of pseudogene-associated lncRNA genes is invariant at stricter expression level cutoffs (Figure S50–53). Altogether, these results ensure the reliability of enhanced functional attributes of pseudogene-associated lncRNA genes.

Both PAS and PAA lncRNA genes show an advantage in serving as regulators of gene expression over NPA lncRNA genes. However, PAS and PAA lncRNA genes exhibit divergent functional advantages. Previous studies [62] have implied that PAS lncRNA genes may inherit the RBP binding sites from parent genes of pseudogenes, and our analysis confirmed this view. Therefore, PAS lncRNA genes have greater protein-binding proficiency compared to PAA lncRNA genes. Conversely, PAA lncRNA genes have an advantage in binding to the mRNAs of parent genes of pseudogenes [65,66].

Furthermore, a transcriptome profiling across nine human tissues from young and old age groups at the single-cell level revealed that pseudogene-associated lncRNA genes exhibited prominent and widespread aging-related expression alterations. These pseudogene-associated lncRNA genes are regulated by aging-related TFs and regulate

580 aging-related pathways. Their close association with aging was further confirmed to
581 be conserved in mouse by RNA-seq data and RT-qPCR. Taken together,
582 pseudogene-associated lncRNA genes will constitute a significant resource for the
583 mining of functional lncRNA genes and provide informative insights into human
584 aging biology in the future.
585

586 ## Materials and methods

587 ### Identification of Pseudogenes

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

598 ### Construction of lncRNA Annotation

599 We complemented the Ensemble lncRNA annotation with long-read transcriptome
600 sequencing data. We first processed mouse, chicken, and zebrafish Iso-seq data
601 generated in our lab. The mouse data have been published [14] and deposited in the
602 GEO database as GSE176018. GSE176018 contains RNA-seq and Iso-seq data for six
603 tissues (including brain, cerebellum, heart, colon, ovary, and testis) from adult
604 C57BL/6J female and male mice. Chicken and zebrafish data include RNA-seq and
605 Iso-seq data for six tissues (including brain, cerebellum, heart, colon, ovary, and testis)
606 from adult female and male chickens and five tissues (including brain, heart, colon,
607 ovary, and testis) from adult female and male zebrafish. Raw Iso-seq data were

processed by CCS (v6.2.0) to generate circular consensus sequences (parameters: -j 64 --min-passes 1 --min-length 300 --min-rq 0.9). Circular consensus sequences were processed by lima (v2.3.99) to remove primer sequences to produce full-length sequences (parameters: -j 64 --isoseq). Full-length non-chimeric sequences were generated by Isoseq3 refine (v3.5.0) (parameters: -j 64 --require-polya). Full-length transcript sequences were generated by Isoseq3 cluster (v3.5.0). RNA-seq data were used by LoRDEC (v0.9) to correct full-length transcript sequences [67]. Corrected transcript sequences longer than 200 were compared to the genome using Minimap2 (v2.26-r1175) (parameters: -t 64 -ax splice --secondary=no -C5 -O6,24 -B4 -uf) [68]. Redundant isoforms were collapsed using the script collapse_isoforms_by_sam.py from cDNA_Cupcake (v29.0.0; https://github.com/Magdoll/cDNA_Cupcake) (parameters: --dun-merge-5- shorter -c 0.85 -i 0.95 --cpus 28). 5' degraded transcripts were removed using get_abundance_post_collapse.py and filter_away_subset.py. The above process completed the transcriptome annotation based on long-read transcriptome sequencing data in mouse, chicken, and zebrafish. We obtained human transcriptome annotations based on long-read transcriptome sequencing data from GTEx V9 release.

Bedtools (v2.30.0) was used to select novel genes that did not overlap with Ensemble genes other than pseudogenes from the above annotations based on long-read transcriptome sequencing. We then removed all genes with evidence of coding potential. CPC2 (v0.1) [69], CNCI (v2) [70], PLEK (v1.2) [71], and LGC (v1.0) [72] were used to detect protein-coding potential. InterProScan (v5.61-93.0) [73] was used to detect transcripts with the Pfam protein domains. Only genes that successfully passed all five filters for all their transcripts were identified as novel lncRNA genes based on long-read transcriptome sequencing.

Next, we obtained high-quality lncRNA gene datasets annotated with short-read transcriptome sequencing data based on multiple developmental stages in seven tissues for seven species (human, rhesus macaque, mouse, rat, rabbit, opossum, and chicken) from the previous study by Kaessmann et al [1]. We also obtained a

637 high-quality lncRNA gene dataset from the FishGET database for zebrafish, annotated
638 based on short-read transcriptome sequencing data [74]. These lncRNA gene dataset
639 were mapped to the genome used in this study using liftOver, and transcripts with
640 changes in transcript structure after mapping were excluded. lncRNAs that overlapped
641 with Ensemble genes other than pseudogenes and novel lncRNAs annotated based on
642 long-read transcriptome sequencing were also eliminated. The remaining lncRNAs
643 were then used as novel lncRNAs annotated based on short-read transcriptome
644 sequencing. These novel lncRNA annotations were merged with the Ensembl
645 annotations as complements.

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

647 We identified two categories of pseudogene-associated lncRNA genes based on two
648 different overlap patterns. Among the total lncRNA genes, we considered genes
649 whose exons overlap with pseudogenes on the same strand as PAS lncRNA genes, and
650 genes whose exons overlap with pseudogenes on the opposite strand as PAA lncRNA
651 genes. All remaining lncRNA genes were classified as NPA lncRNA genes. The data
652 for the two categories of pseudogene-associated lncRNA genes are provided in Table
653 S1 and Table S2.

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

655 To evaluate the enrichment of lncRNA genes at pseudogenes, Bedtools shuffle
656 (v2.30.0) was used to randomly generate 1000 sets of shuffled regions with the same
657 number and equal lengths as pseudogenes. Shuffled regions were not to be placed in
658 exons of protein-coding genes. Bedtools intersect was used to screen for overlap
659 between pseudogenes or shuffled regions and exons of lncRNA genes. Fold
660 enrichment is a ratio of an observed overlap compared to the expected overlap based
661 on shuffling. The number of pseudogenes overlapping exons of lncRNA genes was
662 calculated as the observed overlap. The mean of the number of shuffled regions
663 overlapping exons of lncRNA genes in each set of shuffled regions was calculated as
664 the expected overlap. The empirical p-value of enrichment is defined as the number of
665 shuffled region sets that show an equal or greater overlap than the observed overlap.

To exclude the potential effects of chromosomal biases, we utilized the Bedtools shuffle parameter "-chrom" to make the shuffled regions exhibit the same chromosomal distribution as the pseudogenes. To exclude the effects of repetitive sequences, we downloaded annotations of repetitive sequences from UCSC Genome Browser and excluded pseudogenes and lncRNA genes overlapping with repetitive sequences in the enrichment analysis. We utilized the Bedtools shuffle parameter "-excl" to ensure that the shuffled regions are not placed in repetitive sequences.

Identification of Homologous lncRNA Genes and Estimation of lncRNA Gene Age

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

Sequence Conservation and Variation Analysis

Genome-wide nucleotide resolution PhastCons scores were downloaded from the UCSC Genome Browser. Average PhastCons scores for exonic regions were calculated using bigWigAverageOverBed. Human common (minor allele frequency (MAF) $\geq$ 0.01) single nucleotide polymorphism (SNP) data were downloaded from the dsSNP database. The number of SNPs in exon regions was obtained by Bedtools

694 intersect. We obtained the derived allele frequencies of SNPs from phase 3 of the
695 1000 Genomes Project. To avoid the bias caused by the high mutation rate in the CpG
696 island region on the analysis, we removed the CpG island region in the analysis of
697 SNP derived allele frequency and density. The CADD scores of the SNPs were
698 obtained from the CADD database (https://cadd.gs.washington.edu/).

699 **Expression Analysis**

700 We downloaded RNA-seq data from E-MTAB-2836 for a total of 200 samples from
701 32 different human tissues. C57BL/6J male and female adult mouse multiple tissues
702 (including brain, cerebellum, heart, colon, and gonad) RNA-seq data were obtained
703 from a previous study in our laboratory and stored in GSE176018. Raw FASTQ files
704 were quality controlled using FastQC (v0.11.9) and Trim Galore (v0.6.7), and reads
705 were aligned to the reference genome using STAR (v2.7.10b) [75] (parameters:
706 --runMode alignReads --runThreadN 28 --outSAMtype BAM SortedByCoordinate
707 --readFilesCommand zcat --sjdbOverhang 149 --twopassMode Basic
708 --outBAMsortingThreadN 20 --outSAMmultNmax 1 --outSJfilterReads Unique
709 --outSAMattrIHstart 0 --outSAMstrandField intronMotif). Reads with mapping scores
710 (MAPQ) < 20 were filtered out using SAMtools view (v1.16.1) [76]. Finally, gene
711 expression was quantified using featureCounts (v2.0.3) [77] and gene expression was
712 quantified using transcripts per kilobase million (TPM) to normalize gene expression
713 levels. Genes expressed in any sample (TPM ≥ 1) were used for subsequent analysis.

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

715 We obtained human epigenomic data from the NIH Roadmap Epigenomics Mapping
716 Consortium (http://egg2.wustl.edu/roadmap/). We downloaded DHS data for 53
717 consolidated epigenomes, H3K27ac peak data for 98 consolidated epigenomes,
718 H3K9ac peak data for 62 consolidated epigenomes, H3K4me2 peak data for 24
719 consolidated epigenomes, H3K4me3 peak data for 127 consolidated epigenomes,
720 fractional methylation data from WGBS for 37 consolidated epigenomes, and
721 chromHMM chromatin status for 127 consolidated epigenomes. In addition, we
722 obtained human housekeeping consensus peaks from a study based on the analysis of

723 1785 ATAC-seq and 231 scATAC-seq datasets [78]. We calculated the distribution of
724 housekeeping consensus peaks from ATAC-seq, DHS (DNase I hypersensitivity sites),
725 and four histone modification peaks in a total of 4kb regions upstream and
726 downstream of the gene TSSs. The distribution of each peak relative to the
727 transcription start site (TSS) was generated by deepTools [79], and the average
728 fractional methylation of the region around the TSS was calculated by
729 bigWigAverageOverBed. Genome-wide binding sites for TFs based on ChIP-seq data
730 were downloaded from the GTRD database [80]. We identified the proximal promoter
731 of each gene (±500 bp from the TSS) according to the criteria of the FANTOM5
732 consortium. Among the proximal promoters with active promoter states, we counted
733 the ratio of active promoter chromHMM state regions.

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

735 N6-methyladenosine (m6A) site data based on m6A-seq and MeRIP-seq data were
736 downloaded from the REPIC database [81]. Transcriptome-wide binding site data for
737 a large number of RNA-binding proteins based on CLIP-seq were downloaded from
738 the POSTAR3 database [82]. We determined the binding ratio (binding length/total
739 exon length) of each RBP on each gene, and RBP richness was defined as the number
740 of types of different RBPs. The Shannon diversity index and Bray-Curtis similarity of
741 RBPs were all calculated using the R package vegan (v2.6-2). In order to evaluate
742 RBP similarity between each category of pseudogene-associated lncRNAs and
743 transcripts from parent genes of pseudogenes, two types of gene pairs were utilized as
744 controls: pseudogene-associated lncRNA genes and random protein-coding genes
745 exhibiting a matched expression level with parent genes of pseudogenes, as well as
746 parent genes of pseudogenes and random lncRNA genes exhibiting a matched
747 expression level with pseudogene-associated lncRNA genes.

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

749 We calculated Spearman expression correlations and adjusted p-values (Holm's
750 multiple comparison test) between each lncRNA gene and each protein-coding gene.
751 To reduce false positives, we downloaded RNA-seq data from E-MTAB-6814 for a

752 total of 313 samples from 7 tissues and 23 developmental stages in human and from
753 E-MTAB-6798 for a total of 317 samples from 7 tissues and 14 developmental stages
754 in mouse. These data, along with those previously used for expression analysis, were
755 used for screening of co-expressed genes for lncRNA genes. Given the accessible
756 chromatin context and the disproportionate RNA expression observed in adult testis,
757 testis samples were excluded from this analysis. We identified co-expressed gene
758 pairs as those with correlation coefficients greater than 0.75 and adjusted p-values less
759 than 0.001. We further analyzed the enrichment of co-expressed genes for each
760 category of lncRNA genes using all lncRNA genes as background. The co-expressed
761 genes significantly enriched (two-sided Fisher's exact test, $P < 0.05$) by each category
762 of lncRNA genes were analyzed for the enrichment of GO entries. GO enrichment
763 analysis using protein-coding genes co-expressed with at least one lncRNA gene as
764 background. Enrichment analysis was performed using clusterProfiler (v4.8.1) [83].

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

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

771 The filtered UMI count matrices were imported into R using the Read10X
772 function and then analyzed using the R package Seurat (v4.9.9). After merging Seurat
773 objects for samples from the same tissue, cells with > 500 expressed genes and < 50%
774 of mitochondrial reads were retained. The NormalizeData function was used to
775 normalize the data. The identification of variable genes was conducted using the
776 FindVariableFeatures function, resulting in the selection of 5000 genes for subsequent
777 analysis. Principal component analysis (PCA) was performed using the RunPCA
778 function, and the RunHarmony function was used to remove batch effects from the
779 data. The top 20 PCs were used for clustering and U-MAP analysis. The
780 FindAllMarkers function was used to calculate the marker genes for each cluster. Cell

781 types were identified based on the expression levels of canonical marker genes (Table
782 S4). Genes expressed in each cell type (positive cell rate ≥ 0.01) were used for
783 subsequent analysis.

784 We performed differential expression analysis between young and old groups
785 using the FindMarkers function in Seurat. Oocytes are not subjected to differential
786 expression analysis, as they are essentially absent from old ovaries. Genes with
787 |Log2FC| ≥ 0.25 and adjusted P-values < 0.05 were defined as aging-related DEGs.
788 Aging-related DEGs in each category of lncRNA genes were analyzed for GO
789 enrichment analysis. We used the guilt-by-association method to associate the
790 lncRNA genes with GO entries based on their co-expressed protein-coding genes.

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

792 We downloaded RNA-seq data from GSE225576 for a total of 96 samples from 8
793 tissues of 6-, 24-, and 30-months old male C57BL/6J mice. These RNA-seq data were
794 processed as before. Differential expression analysis was performed between 6- and
795 24- or 30-month-old mice using DESeq2 (v.1.36.0) [84].

796 **Mice and Tissue Collection**

797 All procedures complied with the guidelines approved by the Animal Experimentation
798 Ethics Committee of Huazhong Agricultural University. Young (3-month-old) female
799 C57BL/6 mice were purchased from Wuhan Chunyuhong Experimental Animal Feed
800 corporation limited company, and aged (18-month-old) female C57BL/6 mice were
801 obtained from Wuhan Shubeili Biotechnology corporation limited company. The mice
802 were euthanized by cervical dislocation. Tissue samples were then collected from
803 brain, heart, kidney, liver, and lung. The collected tissues were immediately frozen in
804 liquid nitrogen.

805 **Quantitative Reverse Transcription PCR**

806 Total RNA was extracted from mouse tissues using the FastPure Cell/Tissue Total
807 RNA Isolation Kit V2 (Catalog No. RC112-01, Vazyme, Nanjing, China). Reverse
808 transcription was performed using the HiScript II Q RT SuperMix for qPCR (+gDNA

809 wiper) (Catalog No. R223-01, Vazyme, Nanjing, China). Quantitative real-time PCR
810 was carried out using the ChamQ Universal SYBR qPCR Master Mix (Catalog No.
811 Q711-02, Vazyme, Nanjing, China). Relative RNA expression was calculated using
812 the Ct ($2^{-\Delta\Delta Ct}$) method. The RT-qPCR primers are listed in Table S7.
813

814 Ethical statement

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

820 Data availability

821 The raw sequence data reported in this paper have been deposited in the Genome
822 Sequence Archive [85] in National Genomics Data Center [86], China National
823 Center for Bioinformation / Beijing Institute of Genomics, Chinese Academy of
824 Sciences (GSA: CRA023933, reviewer link:
825 https://ngdc.cncb.ac.cn/gsa/s/QV6bw2mz; BioProject: PRJCA037421) that are
826 publicly accessible at https://ngdc.cncb.ac.cn/gsa. The genomic annotations
827 containing all pseudogene-associated lncRNA genes and lncRNA gene classification
828 information used in this study have been deposited in the OMIX, China National
829 Center for Bioinformation / Beijing Institute of Genomics, Chinese Academy of
830 Sciences (https://ngdc.cncb.ac.cn/omix: accession no. OMIX009515, reviewer link:
831 https://ngdc.cncb.ac.cn/omix/preview/CNojc0bX).

832 Code availability

833 All scripts used in this manuscript have been deposited in Github
834 (https://github.com/Z-H-Zhang/Pseudogene-associated_lncRNA_gene) and BioCode
835 (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. **Hao Zuo:** Formal analysis, Visualization. **Lu Chen:** Methodology. **Meng-Wei Shi:** Software, Visualization. **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)
0009-0004-0031-6256 (Hao Zuo)
0000-0003-4031-5333 (Lu Chen)
0000-0001-6027-9742 (Meng-Wei Shi)
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] Liu SJ, Horlbeck MA, Cho SW, Birk HS, Malatesta M, He D, et al. CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. Science 2017;355.

[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 analysis of pseudogenes across three phyla. PNAS 2014;111:13361–6.

[9] Echols N, Harrison P, Balasubramanian S, Luscombe NM, Bertone P, Zhang Z, et al. Comprehensive analysis of amino acid and nucleotide composition in eukaryotic genomes, comparing genes and pseudogenes. Nucleic Acids Res 2002;30:2515–23.

[10] Cheetham SW, Faulkner GJ, Dinger ME. Overcoming challenges and dogmas to understand the functions of pseudogenes. Nat Rev Genet 2020;21:191–201.

[11] Zhang Z, Harrison PM, Liu Y, Gerstein M. Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. Genome Res 2003;13:2541–58.

[12] Torrents D, Suyama M, Zdobnov E, Bork P. A genome-wide survey of human pseudogenes. Genome Res 2003;13:2559–67.

[13] Harrison PM, Hegyi H, Balasubramanian S, Luscombe NM, Bertone P, Echols N, et al. Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. Genome Res 2002;12:272–80.

[14] Qian SH, Chen L, Xiong Y-L, Chen Z-X. Evolution and function of developmentally dynamic pseudogenes in mammals. Genome Biol 2022;23:235.

[15] Rapicavoli NA, Qu K, Zhang J, Mikhail M, Laberge R-M, Chang HY. A mammalian pseudogene lncRNA at the interface of inflammation and anti-inflammatory therapeutics. eLife 2013;2:e00762.

[16] Scarola M, Comisso E, Pascolo R, Chiaradia R, Maria Marion R, Schneider C, et al. Epigenetic silencing of Oct4 by a complex containing SUV39H1 and Oct4 pseudogene lncRNA. Nat Commun 2015;6:7631.

[17] Liu J, Liu Z-X, Wu Q-N, Lu Y-X, Wong C-W, Miao L, et al. Long noncoding RNA AGPG regulates PFKFB3-mediated tumor glycolytic reprogramming. Nat Commun 2020;11:1507.

[18] Feng J, Yang G, Liu Y, Gao Y, Zhao M, Bu Y, et al. LncRNA PCNAP1 modulates hepatitis B virus replication and enhances tumor growth of liver cancer. Theranostics 2019;9:5227–45.

[19] Lai Y, Li J, Zhong L, He X, Si X, Sun Y, et al. The pseudogene PTENP1 regulates smooth muscle cells as a competing endogenous RNA. Clin Sci (Lond) 2019;133:1439–55.

[20] Zhang R, Kuo R, Coulter M, Calixto CPG, Entizne JC, Guo W, et al. A high-resolution single-molecule sequencing-based Arabidopsis transcriptome using novel methods of Iso-seq analysis. Genome Biol 2022;23:149.

[21] Wang K, Wang D, Zheng X, Qin A, Zhou J, Guo B, et al. Multi-strategic RNA-seq

922 analysis reveals a high-resolution transcriptional landscape in cotton. Nat
923 Commun 2019;10:4714.
924 [22] Zhang Z, Carriero N, Zheng D, Karro J, Harrison PM, Gerstein M. PseudoPipe: an
925 automated pseudogene identification pipeline. Bioinformatics 2006;22:1437–9.
926 [23] Sisu C, Muir P, Frankish A, Fiddes I, Diekhans M, Thybert D, et al. Transcriptional
927 activity and strain-specific history of mouse pseudogenes. Nat Commun
928 2020;11:3695.
929 [24] Lam HYK, Khurana E, Fang G, Cayting P, Carriero N, Cheung K-H, et al.
930 Pseudofam: the pseudogene families database. Nucleic Acids Res 2008;37:D738–
931 D43.
932 [25] Hillier LW, Miller W, Birney E, Warren W, Hardison RC, Ponting CP, et al.
933 Sequence and comparative analysis of the chicken genome provide unique
934 perspectives on vertebrate evolution. Nature 2004;432:695–716.
935 [26] Lee S-H, Eldi P, Cho S-Y, Rangasamy D. Control of chicken CR1 retrotransposons
936 is independent of Dicer-mediated RNA interference pathway. BMC Biol
937 2009;7:53.
938 [27] Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, et al. The
939 zebrafish reference genome sequence and its relationship to the human genome.
940 Nature 2013;496:498–503.
941 [28] Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, et al. The
942 evolution of lncRNA repertoires and expression patterns in tetrapods. Nature
943 2014;505:635–40.
944 [29] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP:
945 the NCBI database of genetic variation. Nucleic Acids Res 2001;29:308–11.
946 [30] Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general
947 framework for estimating the relative pathogenicity of human genetic variants.
948 Nat Genet 2014;46:310–5.
949 [31] Cai D, Han JJ. Aging-associated lncRNAs are evolutionarily conserved and
950 participate in NFκB signaling. Nat Aging 2021;1:438–53.

951 [32] Guttman M, Rinn JL. Modular regulatory principles of large non-coding RNAs.
952 Nature 2012;482:339–46.
953 [33] Jiang S, Cheng SJ, Ren LC, Wang Q, Kang YJ, Ding Y, et al. An expanded
954 landscape of human long noncoding RNA. Nucleic Acids Res 2019;47:7842–56.
955 [34] de Goede OM, Nachun DC, Ferraro NM, Gloudemans MJ, Rao AS, Smail C, et al.
956 Population-scale tissue transcriptomics maps long non-coding RNAs to complex
957 disease. Cell 2021;184:2633–48.e19.
958 [35] Schuettengruber B, Martinez A-M, Iovino N, Cavalli G. Trithorax group proteins:
959 switching genes on and keeping them active. Nat Rev Mol Cell Biol 2011;12:799–
960 814.
961 [36] Wang M, Chen Z, Zhang Y. CBP/p300 and HDAC activities regulate H3K27
962 acetylation dynamics and zygotic genome activation in mouse preimplantation
963 embryos. EMBO J 2022;41:e112012.
964 [37] Jin Q, Yu LR, Wang L, Zhang Z, Kasper LH, Lee JE, et al. Distinct roles of
965 GCN5/PCAF-mediated H3K9ac and CBP/p300-mediated H3K18/27ac in nuclear
966 receptor transactivation. EMBO J 2011;30:249–62.
967 [38] Ernst J, Kellis M. Chromatin-state discovery and genome annotation with
968 ChromHMM. Nat Protoc 2017;12:2478–92.
969 [39] Statello L, Guo C-J, Chen L-L, Huarte M. Gene regulation by long non-coding
970 RNAs and its biological functions. Nat Rev Mol Cell Biol 2021;22:96–118.
971 [40] Hentze MW, Castello A, Schwarzl T, Preiss T. A brave new world of RNA-binding
972 proteins. Nat Rev Mol Cell Biol 2018;19:327–41.
973 [41] Huang W, Xiong T, Zhao Y, Heng J, Han G, Wang P, et al. Computational
974 prediction and experimental validation identify functionally conserved lncRNAs
975 from zebrafish to human. Nat Genet 2024;56:124–35.
976 [42] Lee S, Kopp F, Chang T-C, Sataluri A, Chen B, Sivakumar S, et al. Noncoding
977 RNA NORAD Regulates Genomic Stability by Sequestering PUMILIO Proteins.
978 Cell 2016;164:69–80.
979 [43] Hosono Y, Niknafs YS, Prensner JR, Iyer MK, Dhanasekaran SM, Mehra R, et al.

980 Oncogenic Role of THOR, a Conserved Cancer/Testis Long Non-coding RNA.
981 Cell 2017;171:1559–72.e20.
982 [44] Wu G, Su J, Zeng L, Deng S, Huang X, Ye Y, et al. LncRNA BCAN-AS1 stabilizes
983 c-Myc via N6-methyladenosine-mediated binding with SNIP1 to promote
984 pancreatic cancer. Cell Death Differ 2023;30:2213–30.
985 [45] Liu Y, Yang D, Liu T, Chen J, Yu J, Yi P. N6-methyladenosine-mediated gene
986 regulation and therapeutic implications. Trends Mol Med 2023;29:454–67.
987 [46] Gong Y, Liang Y, Liu J, Wei J, Zhang S, Chen F, et al. DDX24 Is Essential for Cell
988 Cycle Regulation in Vascular Smooth Muscle Cells During Vascular Development
989 via Binding to FANCA mRNA. Arterioscler, Thromb, Vasc Biol 2023;43:1653–
990 67.
991 [47] Randolph ME, Afifi M, Gorthi A, Weil R, Wilky BA, Weinreb J, et al. RNA
992 helicase DDX3 regulates RAD51 localization and DNA damage repair in Ewing
993 sarcoma. iScience 2024;27:108925.
994 [48] Riva V, Garbelli A, Casiraghi F, Arena F, Trivisani Claudia I, Gagliardi A, et al.
995 Novel alternative ribonucleotide excision repair pathways in human cells by
996 DDX3X and specialized DNA polymerases. Nucleic Acids Res 2020;48:11551–
997 65.
998 [49] Gu J, Chen J, Yin Q, Dong M, Zhang Y, Chen M, et al. lncRNA JPX-Enriched
999 Chromatin Microenvironment Mediates Vascular Smooth Muscle Cell
1000 Senescence and Promotes Atherosclerosis. Arterioscler, Thromb, Vasc Biol
1001 2024;44:156–76.
1002 [50] Ruvkun G, Lehrbach N. Regulation and Functions of the ER-Associated Nrf1
1003 Transcription Factor. Cold Spring Harb Perspect Biol 2023;15.
1004 [51] Suryadevara V, Hudgins AD, Rajesh A, Pappalardo A, Karpova A, Dey AK, et al.
1005 SenNet recommendations for detecting senescent cells in different tissues. Nat
1006 Rev Mol Cell Biol 2024.
1007 [52] Vernier M, Bourdeau V, Gaumont-Leclerc MF, Moiseeva O, Bégin V, Saad F, et al.
1008 Regulation of E2Fs and senescence by PML nuclear bodies. Genes Dev

1009 2011;25:41–50.
1010 [53] Yang S, Liu C, Jiang M, Liu X, Geng L, Zhang Y, et al. A single-nucleus
1011 transcriptomic atlas of primate liver aging uncovers the pro-senescence role of
1012 SREBP2 in hepatocytes. Protein Cell 2023;15:98–120.
1013 [54] McIntyre RL, Liu YJ, Hu M, Morris BJ, Willcox BJ, Donlon TA, et al.
1014 Pharmaceutical and nutraceutical activation of FOXO3 for healthy longevity.
1015 Ageing Research Reviews 2022;78:101621.
1016 [55] Liu G, Chen H, Liu H, Zhang W, Zhou J. Emerging roles of SIRT6 in human
1017 diseases and its modulators. Med Res Rev 2021;41:1089–137.
1018 [56] Jing Y, Jiang X, Ji Q, Wu Z, Wang W, Liu Z, et al. Genome-wide CRISPR
1019 activation screening in senescent cells reveals SOX5 as a driver and therapeutic
1020 target of rejuvenation. Cell Stem Cell 2023;30:1452–71.e10.
1021 [57] Johnson SC, Rabinovitch PS, Kaeberlein M. mTOR is a key modulator of ageing
1022 and age-related disease. Nature 2013;493:338–45.
1023 [58] Bou Sleiman M, Jha P, Houtkooper R, Williams RW, Wang X, Auwerx J. The
1024 Gene-Regulatory Footprint of Aging Highlights Conserved Central Regulators.
1025 Cell Rep 2020;32:108203.
1026 [59] Tang H, Xiong Q, Yin M, Feng H, Yao F, Xiao X, et al. LncRNA PVT1 delays skin
1027 photoaging by sequestering miR-551b-3p to release AQP3 expression via ceRNA
1028 mechanism. Apoptosis 2023;28:912–24.
1029 [60] Du H, Zhang H, Yang R, Qiao L, Shao H, Zhang X. Small interfering
1030 RNA-induced silencing lncRNA PVT1 inhibits atherosclerosis via inactivating the
1031 MAPK/NF-κB pathway. Aging 2021;13:24449–63.
1032 [61] Palazzo AF, Koonin EV. Functional Long Non-coding RNAs Evolve from Junk
1033 Transcripts. Cell 2020;183:1151–61.
1034 [62] Ma Y, Liu S, Gao J, Chen C, Zhang X, Yuan H, et al. Genome-wide analysis of
1035 pseudogenes reveals HBBP1's human-specific essentiality in erythropoiesis and
1036 implication in β-thalassemia. Dev Cell 2021;56:478–93.e11.
1037 [63] Li Z, Liu L, Feng C, Qin Y, Xiao J, Zhang Z, et al. LncBook 2.0: integrating human

1038 long non-coding RNAs with multi-omics annotations. Nucleic Acids Res
1039 2022;51:D186–D91.
1040 [64] Kaessmann H. Origins, evolution, and phenotypic impact of new genes. Genome
1041 Res 2010;20:1313–26.
1042 [65] Lister NC, Johnsson P, Waters PD, Morris KV. Pseudogenes: A Novel Source of
1043 Trans-Acting Antisense RNAsAntisense RNAs. In: Poliseno L. (ed) Pseudogenes:
1044 Functions and Protocols. New York, NY: Springer US, 2021, 219–36.
1045 [66] Milligan MJ, Lipovich L. Pseudogene-derived lncRNAs: emerging regulators of
1046 gene expression. Front Genet 2015;5.
1047 [67] Salmela L, Rivals E. LoRDEC: accurate and efficient long read error correction.
1048 Bioinformatics 2014;30:3506–14.
1049 [68] Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics
1050 2018;34:3094–100.
1051 [69] Kang YJ, Yang DC, Kong L, Hou M, Meng YQ, Wei L, et al. CPC2: a fast and
1052 accurate coding potential calculator based on sequence intrinsic features. Nucleic
1053 Acids Res 2017;45:W12–w6.
1054 [70] Sun L, Luo H, Bu D, Zhao G, Yu K, Zhang C, et al. Utilizing sequence intrinsic
1055 composition to classify protein-coding and long non-coding transcripts. Nucleic
1056 Acids Res 2013;41:e166.
1057 [71] Li A, Zhang J, Zhou Z. PLEK: a tool for predicting long non-coding RNAs and
1058 messenger RNAs based on an improved k-mer scheme. BMC Bioinf 2014;15:311.
1059 [72] Wang G, Yin H, Li B, Yu C, Wang F, Xu X, et al. Characterization and
1060 identification of long non-coding RNAs based on feature relationship.
1061 Bioinformatics 2019;35:2949–56.
1062 [73] Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, et al.
1063 InterProScan: protein domains identifier. Nucleic Acids Res 2005;33:W116–20.
1064 [74] Guo C, Duan Y, Ye W, Zhang W, Cheng Y, Shi M, et al. FishGET: A fish gene
1065 expression and transcriptome database with improved accuracy and visualization.
1066 iScience 2023;26:106539.

1067 [75] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR:
1068 ultrafast universal RNA-seq aligner. Bioinformatics 2013;29:15–21.
1069 [76] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence
1070 Alignment/Map format and SAMtools. Bioinformatics 2009;25:2078–9.
1071 [77] Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for
1072 assigning sequence reads to genomic features. Bioinformatics 2013;30:923–30.
1073 [78] Qiuchen M, Xinze W, Chen L, Jiaqi L, Xi X, Sijie C, et al. The full set of potential
1074 open regions (PORs) in the human genome defined by consensus peaks of
1075 ATAC-seq data. bioRxiv 2023:2023.05.30.542889.
1076 [79] Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, et al.
1077 deepTools2: a next generation web server for deep-sequencing data analysis.
1078 Nucleic Acids Res 2016;44:W160–5.
1079 [80] Yevshin I, Sharipov R, Kolmykov S, Kondrakhin Y, Kolpakov F. GTRD: a
1080 database on gene transcription regulation-2019 update. Nucleic Acids Res
1081 2019;47:D100–d5.
1082 [81] Liu S, Zhu A, He C, Chen M. REPIC: a database for exploring the
1083 N6-methyladenosine methylome. Genome Biol 2020;21:100.
1084 [82] Zhao W, Zhang S, Zhu Y, Xi X, Bao P, Ma Z, et al. POSTAR3: an updated platform
1085 for exploring post-transcriptional regulation coordinated by RNA-binding
1086 proteins. Nucleic Acids Res 2022;50:D287–d94.
1087 [83] Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: A universal
1088 enrichment tool for interpreting omics data. Innovation (Camb) 2021;2:100141.
1089 [84] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion
1090 for RNA-seq data with DESeq2. Genome Biol 2014;15:550.
1091 [85] Chen T, Chen X, Zhang S, Zhu J, Tang B, Wang A, et al. The Genome Sequence
1092 Archive Family: Toward Explosive Data Growth and Diverse Data Types.
1093 Genomics Proteomics Bioinformatics 2021.
1094 [86] Database Resources of the National Genomics Data Center, China National Center
1095 for Bioinformation in 2024. Nucleic Acids Res 2024;52:D18–d32.

1096

1097 # Figure legends

1098 **Figure 1 Identification of pseudogene-associated lncRNA genes**

1099 **A**. Tissue samples from zebrafish, chicken and mouse. Somatic tissues were pooled
1100 into one sample for Iso-seq library preparation for each sex. **B**. Length of transcripts
1101 in Ensembl annotation and PacBio Iso-seq. **C**. Exon number of transcripts in Ensembl
1102 annotation and PacBio Iso-seq. **D**. Number of lncRNA annotations in each species. **E**.
1103 Number of pseudogenes in each species. **F**. Methods to identify
1104 pseudogene-associated lncRNA genes. **G**. Number of pseudogene-associated lncRNA
1105 genes in each species. Bar plot showing the proportion of two categories among
1106 pseudogene-associated lncRNA genes. **H**. Expression patterns of zebrafish novel
1107 lncRNA gene *PB.7298* and mouse novel lncRNA gene *PB.5227* which are detected by
1108 Iso-seq and are identified as pseudogene-associated lncRNA genes. Iso-seq, isoform
1109 sequencing.

1110

1111 **Figure 2 lncRNA genes are enriched at pseudogenes**

1112 **A**. Enrichment of sense and antisense lncRNA genes at pseudogenes in each species.
1113 $*P \leq 0.05$, $**P \leq 0.01$, $***P \leq 0.001$. **B**. Sequence conservation (PhastCons score),
1114 for overlapping and non-overlapping regions within the pseudogenes overlapping
1115 with sense and antisense lncRNA genes in human and mouse. **C**. Genomic profile
1116 depicts a high conservation of regions within the pseudogene HOS_PG_6183
1117 overlapping the *ROCK1P1* exons. The per-base phastCons, and the phyloP
1118 conservation scores are presented.

1119

1120 **Figure 3 Functionality characterization and expression profiles of**
1121 **pseudogene-associated lncRNA genes**

1122 **A**. Distribution of PAS lncRNA genes, PAA lncRNA genes, and NPA lncRNA genes
1123 in the branches of the human phylogenetic tree. Tree tips, lncRNA proportions for
1124 each age group. $*P < 0.05$, $**P < 0.01$, $***P < 0.001$, two-sided Fisher's exact test. **B**.

1125 Enrichment of PAS lncRNA genes, PAA lncRNA genes, and NPA lncRNA genes at
1126 different evolutionary ages in human. A total of 2000 randomly selected NPA lncRNA
1127 genes were utilized as controls. **C**. Exonic sequence conservation (PhastCons score),
1128 for CDS, UTR, PAS lncRNA genes, PAA lncRNA genes, NPA lncRNA genes, and
1129 random intergenic regions in human. **D**. Exon common SNP density of CDS, UTR,
1130 PAS lncRNA genes, PAA lncRNA genes, NPA lncRNA genes, and random intergenic
1131 regions in human. **E**. Transcript lengths of protein-coding genes, PAS lncRNA genes,
1132 PAA lncRNA genes, and NPA lncRNA genes in human. **F**. Exonic sequence
1133 conservation (PhastCons score), for PAS lncRNA genes, PAA lncRNA genes, and
1134 NPA lncRNA genes in each age group in human. **G**. Maximum expression levels of
1135 protein-coding genes, PAS lncRNA genes, PAA lncRNA genes, and NPA lncRNA
1136 genes in human. **H**. Expression proportions of protein-coding genes, PAS lncRNA
1137 genes, PAA lncRNA genes, and NPA lncRNA genes at a series of expression
1138 thresholds in human. **I**. Mean expression proportions of protein-coding genes, PAS
1139 lncRNA genes, PAA lncRNA genes, and NPA lncRNA genes in various tissues in
1140 human. Error bars indicate standard deviation of the proportion of expression. Legend
1141 is the same as in Figure 3H. PAS, pseudogene-associated sense; PAA,
1142 pseudogene-associated antisense; NPA, non-pseudogene-associated; Ma, million
1143 years ago; CDS, coding sequence; UTR, untranslated region; SNP, single nucleotide
1144 polymorphism.
1145

1146 **Figure 4 Epigenomic features of pseudogene-associated lncRNA genes**

1147 **A**. Distribution of housekeeping consensus peaks from ATAC-seq in a total of 4kb
1148 regions upstream and downstream of protein-coding gene, PAS lncRNA genes, PAA
1149 lncRNA genes, and NPA lncRNA genes TSSs, and random intergenic regions. **B**.
1150 Distribution of DHS in a total of 4kb regions upstream and downstream of
1151 protein-coding gene, PAS lncRNA genes, PAA lncRNA genes, and NPA lncRNA
1152 genes TSSs, and random intergenic regions. **C**. Distribution of H3K4me3, H3K4me2,
1153 H3K9ac, and H3K27ac peaks in a total of 4kb regions upstream and downstream of

1154 protein-coding gene, PAS lncRNA genes, PAA lncRNA genes, and NPA lncRNA
1155 genes TSSs, and random intergenic regions. **D**. Distribution of ten chromatin
1156 modifiers binding sites in a total of 2kb regions upstream and downstream of
1157 protein-coding gene, PAS lncRNA genes, PAA lncRNA genes, and NPA lncRNA
1158 genes TSSs, and random intergenic regions. **E**. DNA methylation in each 200bp
1159 window within a total of 4kb regions upstream and downstream of protein-coding
1160 gene, PAS lncRNA genes, PAA lncRNA genes, and NPA lncRNA genes TSSs, and
1161 random intergenic regions. Dotted lines indicate average DNA methylation values
1162 over the entire region. Legend is the same as in Figure 4A. **F**. Ratios of active
1163 promoter chromHMM state regions in the proximal promoters of protein-coding
1164 genes, PAS lncRNA genes, PAA lncRNA genes, NPA lncRNA genes, and random
1165 intergenic regions. TSS, transcription start site; DHS, DNase I hypersensitivity site.
1166

1167 **Figure 5 RNA-protein binding related features and putative functions of**
1168 **pseudogene-associated lncRNA genes**

1169 **A**. Density of m6A sites in protein-coding gene transcripts, PAS lncRNAs, PAA
1170 lncRNAs, and NPA lncRNAs. **B**. Richness of RBPs bound to protein-coding gene
1171 transcripts, PAS lncRNAs, PAA lncRNAs, and NPA lncRNAs. **C**. Bray-Curtis
1172 similarity between RBPs bound to pseudogene-associated lncRNAs and RBPs bound
1173 to transcripts of parent genes of pseudogenes. Two types of gene pairs were utilized as
1174 controls: pseudogene-associated lncRNA genes and random protein-coding genes that
1175 exhibited a matched expression level with parent genes of pseudogenes, as well as
1176 parent genes of pseudogenes and random lncRNA genes that exhibited a matched
1177 expression level with pseudogene-associated lncRNA genes. **D**. Top 10 RBPs
1178 enriched for PAS and PAA lncRNAs. A total of 2000 randomly selected NPA
1179 lncRNAs were used as controls. **E**. GO enrichment analysis to annotate the function
1180 of PAS lncRNA genes in human. The representative GO biological processes are
1181 shown. m6A, N6-methyladenosine; RBP, RNA-binding protein; GO, Gene Ontology.
1182

1183 **Figure 6 Aging-related expression changes of pseudogene-associated lncRNA**
1184 **genes across nine human tissues**
1185 **A**. Proportions of aging-related DEGs for protein-coding genes, PAS lncRNA genes,
1186 PAA lncRNA genes, and NPA lncRNA genes in each tissue. **B**. Odds ratios and
1187 confidence intervals (95%) between the proportions of aging-related DEGs in each
1188 category of lncRNA genes in nine tissues. $*P < 0.05$, $**P < 0.01$, $***P < 0.001$,
1189 two-sided Fisher's exact tests were performed for comparisons between each category
1190 of lncRNA genes. **C**. Among aging-related DEGs, proportions of tissue-shared
1191 aging-related DEGs for protein-coding genes, PAS lncRNA genes, PAA lncRNA
1192 genes, and NPA lncRNA genes. **D**. Number of tissues in which the genes exhibited
1193 aging-related differential expression for protein-coding genes, PAS lncRNA genes,
1194 PAA lncRNA genes, and NPA lncRNA genes in each expression breadth group. The
1195 diamond dots represent the average. $*P < 0.05$, $**P < 0.01$, $***P < 0.001$, two-sided
1196 Wilcoxon rank sum test. **E**. Aging-related differential expression of *JPX* in each cell
1197 type of each tissue. **F**. GO enrichment analysis based on guilt-by-association method
1198 to annotate the function of aging-related DEGs in each category of lncRNA genes.
1199 The representative GO biological processes are shown. A total of 2000 randomly
1200 selected NPA lncRNA genes were used as controls for comparison. DEG,
1201 differentially expressed gene; Mic, microglia; FB, fibroblast; Myeloid, myeloid cell;
1202 HSC, hematopoietic stem cell; LMPP, lympho-myeloid primed progenitor; GMP,
1203 granulocyte-monocyte progenitor; MONO, monocyte; pDC, plasmacytoid dendritic
1204 cell; CLP, common lymphoid progenitor; MEP, megakaryocyte-erythroid progenitor;
1205 Mk, megakaryocyte; Ery, erythroid; Basophil, basophil cell; Prolif, proliferative cell;
1206 Macro, macrophage; GC, granulosa cell; T&S, theca and stroma cell; SMC, smooth
1207 muscle cell; EC, endothelial cell; Late SPC, late primary spermatocyte.
1208
1209 **Figure 7 Aging-related expression changes of pseudogene-associated lncRNA**
1210 **genes in mouse**
1211 **A**. Aging-related expression changes of the top 3 PAS and PAA lncRNA genes

1212 differentially expressed in the most mouse tissues. Statistical significance labels are
1213 based on adjusted p-values from DESeq2, *$P \leq 0.05$, **$P \leq 0.01$, ***$P \leq 0.001$. **B**.
1214 Expression levels of *D17H6S56E-5* in five tissues from 3 month and 18 month old
1215 mice, revealed by RT-qPCR. Expression levels are shown as means. Error bars
1216 indicate the standard deviation of expression levels. 3 month, n = 3; 18 month, n = 3
1217 mice. **$P < 0.01$, ***$P < 0.001$, two-sided Student's T test. **C**. Expression levels of
1218 *Gm36161* in five tissues from 3 month and 18 month old mice, revealed by RT-qPCR.
1219 Expression levels are shown as means. Error bars indicate the standard deviation of
1220 expression levels. 3 month, n = 3; 18 month, n = 3 mice. **$P < 0.01$, ***$P < 0.001$,
1221 two-sided Student's T test. **D**. Orthologous pseudogene-associated lncRNA gene pairs
1222 that exhibited aging-related differential expression in both human and mouse. **E**.
1223 Expression levels of *Mou_XLOC_028665*, *Mou_XLOC_013922*, *Pvt1*, and
1224 *Mou_XLOC_000120* were quantified by RT-qPCR in five tissues from 3 month and
1225 18 month old mice. Expression levels are shown as means. Error bars indicate the
1226 standard deviation of expression levels. 3 month, n = 3; 18 month, n = 3 mice. *$P <$
1227 0.05, **$P < 0.01$, ***$P < 0.001$, two-sided Student's T test. RT-qPCR, quantitative
1228 reverse transcription PCR.
1229
1230

1231 ## Supplementary material

1232 **Figure S1 Pipeline for identification of pseudogenes, integration of lncRNA**
1233 **annotations, and identification of the pseudogene-associated lncRNA gene**
1234

1235 **Figure S2 Transcript lengths of Ensemble lncRNA genes, novel lncRNA genes**
1236 **annotated by long-read transcriptome sequencing, and novel lncRNA genes**
1237 **annotated by short-read transcriptome sequencing**
1238

1239 **Figure S3 Exonic sequence conservation for Ensemble lncRNA genes, novel**
1240 **lncRNA genes annotated by long-read transcriptome sequencing, and novel**

1241 **lncRNA genes annotated by short-read transcriptome sequencing**

1242

1243 **Figure S4 Maximum expression levels of Ensemble lncRNA genes, novel**
1244 **lncRNA genes annotated by long-read transcriptome sequencing, and novel**
1245 **lncRNA genes annotated by short-read transcriptome sequencing**

1246

1247 **Figure S5 Proportion of the three categories of pseudogenes in each species**

1248

1249 **Figure S6 Desnsity of the three categories of pseudogenes in each species**

1250

1251 **Figure S7 Pearson correlation analysis between the number of**
1252 **pseudogene-associated lncRNA genes and the total number of lncRNA genes**

1253

1254 **Figure S8 Bar plot showing the proportion of novel lncRNA genes among**
1255 **pseudogene-associated lncRNA genes**

1256

1257 **Figure S9 Enrichment of sense and antisense lncRNA genes at each category of**
1258 **pseudogenes in each species**

1259

1260 **Figure S10 Distribution of PAS lncRNA genes, PAA lncRNA genes, and NPA**
1261 **lncRNA genes in the branches of the mouse phylogenetic tree**
1262 Tree tips, lncRNA proportions for each age group. *$P < 0.05$, **$P < 0.01$, ***$P <$
1263 0.001, two-sided Fisher's exact test.

1264

1265 **Figure S11 Enrichment of PAS lncRNA genes, PAA lncRNA genes, and NPA**
1266 **lncRNA genes at different evolutionary ages in mouse**
1267 A total of 2000 randomly selected NPA lncRNAs were utilized as controls.

1268

1269 **Figure S12 Exonic sequence conservation (PhastCons score), for CDS, UTR,**

1270 **PAS lncRNA genes, PAA lncRNA genes, NPA lncRNA genes, and random**
1271 **intergenic regions in mouse**
1272
1273 **Figure S13 Exonic sequence conservation (PhastCons score), for PAS lncRNA**
1274 **genes, PAA lncRNA genes, and NPA lncRNA genes in each age group in mouse**
1275
1276 **Figure S14 Mean derived allele frequency of non-CpG SNPs segregating in**
1277 **African populations (1000 Genomes Project)**
1278 Error bars, 95% confidence intervals based on 1000 bootstrap resampling replicates.
1279
1280 **Figure S15 CADD scores for SNPs in CDS, UTR, PAS lncRNA genes, PAA**
1281 **lncRNA genes, NPA lncRNA genes, and random intergenic regions in human**
1282 CADD, Combined Annotation-Dependent Depletion.
1283
1284 **Figure S16 Transcript lengths of protein-coding genes, PAS lncRNA genes, PAA**
1285 **lncRNA genes, and NPA lncRNA genes in mouse**
1286
1287 **Figure S17 Average number of exons in transcripts of protein-coding genes,**
1288 **PAS lncRNA genes, PAA lncRNA genes, and NPA lncRNA genes in human and**
1289 **mouse**
1290 Error bars, 95% confidence intervals based on 1000 bootstrap resampling replicates.
1291
1292 **Figure S18 Maximum expression levels of protein-coding genes, PAS lncRNA**
1293 **genes, PAA lncRNA genes, and NPA lncRNA genes in various tissues in human**
1294
1295 **Figure S19 Maximum expression levels of protein-coding genes, PAS lncRNA**
1296 **genes, PAA lncRNA genes, and NPA lncRNA genes in mouse**
1297
1298 **Figure S20 Expression proportions of protein-coding genes, PAS lncRNA genes,**

1299 **PAA lncRNA genes, and NPA lncRNA genes at a series of expression thresholds**
1300 **in mouse**
1301
1302 **Figure S21 Mean expression proportions of protein-coding genes, PAS lncRNA**
1303 **genes, PAA lncRNA genes, and NPA lncRNA genes in various tissues in mouse**
1304 Error bars indicate standard deviation of the proportion of expression. Legend is the
1305 same as in Figure 3H.
1306
1307 **Figure S22 Maximum expression levels of protein-coding genes, PAS lncRNA**
1308 **genes, PAA lncRNA genes, and NPA lncRNA genes in various tissues in mouse**
1309
1310 **Figure S23 Number of tissues with expression of protein-coding genes, PAS**
1311 **lncRNA genes, PAA lncRNA genes, and NPA lncRNA genes in human and mouse**
1312
1313 **Figure S24 Shannon diversity index of RBPs bound to protein-coding gene**
1314 **transcripts, PAS lncRNAs, PAA lncRNAs, and NPA lncRNAs in human**
1315
1316 **Figure S25 Density of m6A sites in protein-coding gene transcripts, PAS**
1317 **lncRNAs, PAA lncRNAs, and NPA lncRNAs in each expression level in human**
1318
1319 **Figure S26 Richness of RBPs bound to protein-coding gene transcripts, PAS**
1320 **lncRNAs, PAA lncRNAs, and NPA lncRNAs in each expression level in human**
1321
1322 **Figure S27 Shannon diversity index of RBPs bound to protein-coding gene**
1323 **transcripts, PAS lncRNAs, PAA lncRNAs, and NPA lncRNAs in each expression**
1324 **level in human**
1325
1326 **Figure S28 Number of RBPs enriched by PAS lncRNAs and PAA lncRNAs**
1327

1328 **Figure S29 Number of co-expressed genes enriched by each category of lncRNA**
1329 **genes at each p-value level in human**
1330 A total of 2000 randomly selected NPA lncRNA genes were utilized as controls.
1331
1332 **Figure S30 Number of co-expressed genes enriched by each category of lncRNA**
1333 **genes at each p-value level in mouse**
1334 A total of 2000 randomly selected NPA lncRNA genes were utilized as controls.
1335
1336 **Figure S31 GO enrichment analysis to annotate the function of NPA lncRNA**
1337 **genes in mouse**
1338
1339 **Figure S32 GO enrichment analysis to annotate the function of PAS lncRNA**
1340 **genes in mouse**
1341
1342 **Figure S33 GO enrichment analysis to annotate the function of PAA lncRNA**
1343 **genes in mouse**
1344
1345 **Figure S34 GO enrichment analysis to annotate the function of PAA lncRNA**
1346 **genes in human**
1347
1348 **Figure S35 UMAP plots showing the annotated cell types in each tissue**
1349 Cell types are indicated by the different colors.
1350
1351 **Figure S36 Expression proportions for protein-coding genes, PAS lncRNA**
1352 **genes, PAA lncRNA genes, and NPA lncRNA genes in each cell type of each tissue**
1353
1354 **Figure S37 Expression proportions of protein-coding genes, PAS lncRNA genes,**
1355 **PAA lncRNA genes, and NPA lncRNA genes in each tissue**
1356 Legend is the same as in Figure S36.

1357

1358 **Figure S38 Number of cell types with expression of protein-coding genes, PAS**
1359 **lncRNA genes, PAA lncRNA genes, and NPA lncRNA genes**

1360

1361 **Figure S39 Proportions of up-regulated and down-regulated genes for**
1362 **protein-coding genes, PAS lncRNA genes, PAA lncRNA genes, and NPA lncRNA**
1363 **genes in each cell type of each tissue**

1364 Legend is the same as in Figure S36. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, two-sided
1365 Fisher's exact tests were conducted to compare pseudogene-associated lncRNA genes
1366 with NPA lncRNA genes.

1367

1368 **Figure S40 Among aging-related DEGs, proportions of tissue-specific genes**
1369 **(Expressed in only one tissue) for protein-coding genes, PAS lncRNA genes, PAA**
1370 **lncRNA genes, and NPA lncRNA genes**

1371 *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, two-sided Fisher's exact test.

1372

1373 **Figure S41 Enrichment of TF binding sites for each category of lncRNA genes**
1374 **in human**

1375 The representative TFs are shown. A total of 2000 randomly selected NPA lncRNA
1376 genes were utilized as controls. TF, transcription factor.

1377

1378 **Figure S42 Aging-related expression changes of pseudogene-associated lncRNA**
1379 **genes across eight mouse tissues**

1380 **A**. Proportions of aging-related DEGs for protein-coding genes, PAS lncRNA genes,
1381 PAA lncRNA genes, and NPA lncRNA genes in each tissue. **B**. Odds ratios and
1382 confidence intervals (95%) between the proportions of aging-related DEGs in each
1383 category of lncRNA genes in eight tissues. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$,
1384 two-sided Fisher's exact tests were performed for comparisons between each category
1385 of lncRNA genes. **C**. Among aging-related DEGs, proportions of tissue-shared

aging-related DEGs for protein-coding genes, PAS lncRNA genes, PAA lncRNA genes, and NPA lncRNA genes. **D**. Number of tissues in which the genes exhibited aging-related differential expression for protein-coding genes, PAS lncRNA genes, PAA lncRNA genes, and NPA lncRNA genes in each expression breadth group. The diamond dots represent the average. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, two-sided Wilcoxon rank sum test.

Figure S43 Expression levels of *ENSMUSG00000121500*, *ENSMUSG00000121378*, *Mou_XLOC_017668*, and *Mou_XLOC_000586* in five tissues from 3 month and 18 month old mice, revealed by RT-qPCR

Expression levels are shown as means. Error bars indicate standard deviation of the expression levels. 3 month, n = 3; 18 month, n = 3 mice. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, two-sided Student's T test.

Figure S44 Aging-related expression changes for *Mou_XLOC_028665*, *Mou_XLOC_013922*, *Pvt1*, and *Mou_XLOC_000120*, revealed by RNA-seq analysis

Statistical significance labels are based on adjusted p-values from DESeq2, *$P \leq 0.05$, **$P \leq 0.01$, ***$P \leq 0.001$.

Figure S45 Venn plot shows the overlap between LncBook database and our human lncRNA gene annotation

The pie plot on the right shows the sources of lncRNA genes that are only present in our human lncRNA gene annotation.

Figure S46 Functionality characterization of PAS lncRNA genes at a series of overlap length thresholds in human

A. Enrichment of evolutionary ages. **B**. Exonic sequence conservation (PhastCons score). **C**. Transcript lengths. **D**. Maximum expression levels. **E**. Density of m6A sites.

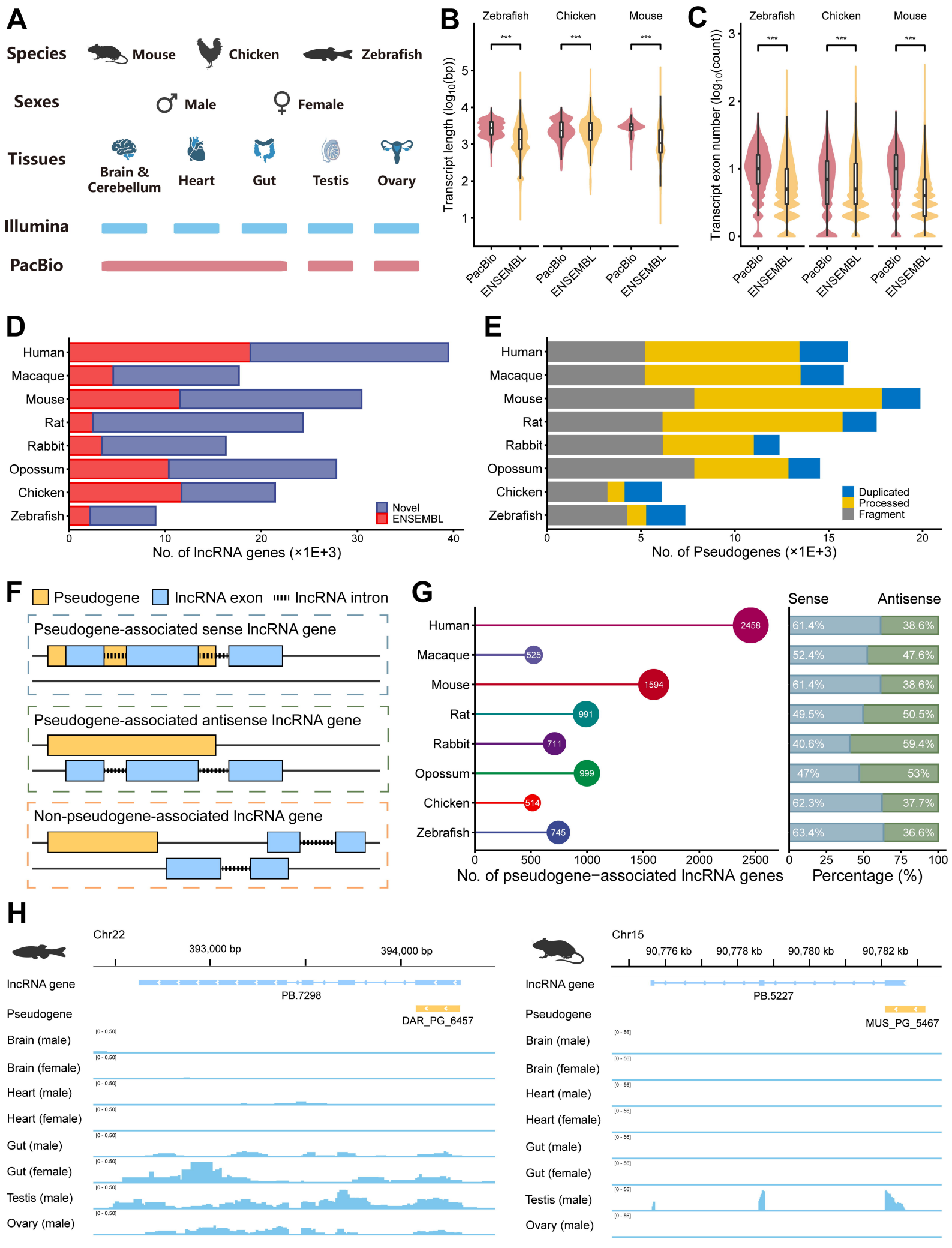
1415 **F**. RBP richness. **G**. RBP diversity. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, Wilcoxon
1416 rank sum test.
1417
1418 **Figure S47 Functionality characterization of PAA lncRNA genes at a series of**
1419 **overlap length thresholds in human**
1420 **A**. Enrichment of evolutionary ages. **B**. Exonic sequence conservation (PhastCons
1421 score). **C**. Transcript lengths. **D**. Maximum expression levels. **E**. Density of m6A sites.
1422 **F**. RBP richness. **G**. RBP diversity. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, Wilcoxon
1423 rank sum test.
1424
1425 **Figure S48 Functionality characterization of PAS lncRNA genes at a series of**
1426 **overlap length thresholds in mouse**
1427 **A**. Enrichment of evolutionary ages. **B**. Exonic sequence conservation (PhastCons
1428 score). **C**. Transcript lengths. **D**. Maximum expression levels. *$P < 0.05$, **$P < 0.01$,
1429 ***$P < 0.001$, Wilcoxon rank sum test.
1430
1431 **Figure S49 Functionality characterization of PAA lncRNA genes at a series of**
1432 **overlap length thresholds in mouse**
1433 **A**. Enrichment of evolutionary ages. **B**. Exonic sequence conservation (PhastCons
1434 score). **C**. Transcript lengths. **D**. Maximum expression levels. *$P < 0.05$, **$P < 0.01$,
1435 ***$P < 0.001$, Wilcoxon rank sum test.
1436
1437 **Figure S50 Functionality characterization of PAS lncRNA genes at a series of**
1438 **expression level thresholds in human**
1439 **A**. Enrichment of evolutionary ages. **B**. Exonic sequence conservation (PhastCons
1440 score). **C**. Transcript lengths.
1441
1442 **Figure S51 Functionality characterization of PAA lncRNA genes at a series of**
1443 **expression level thresholds in human**

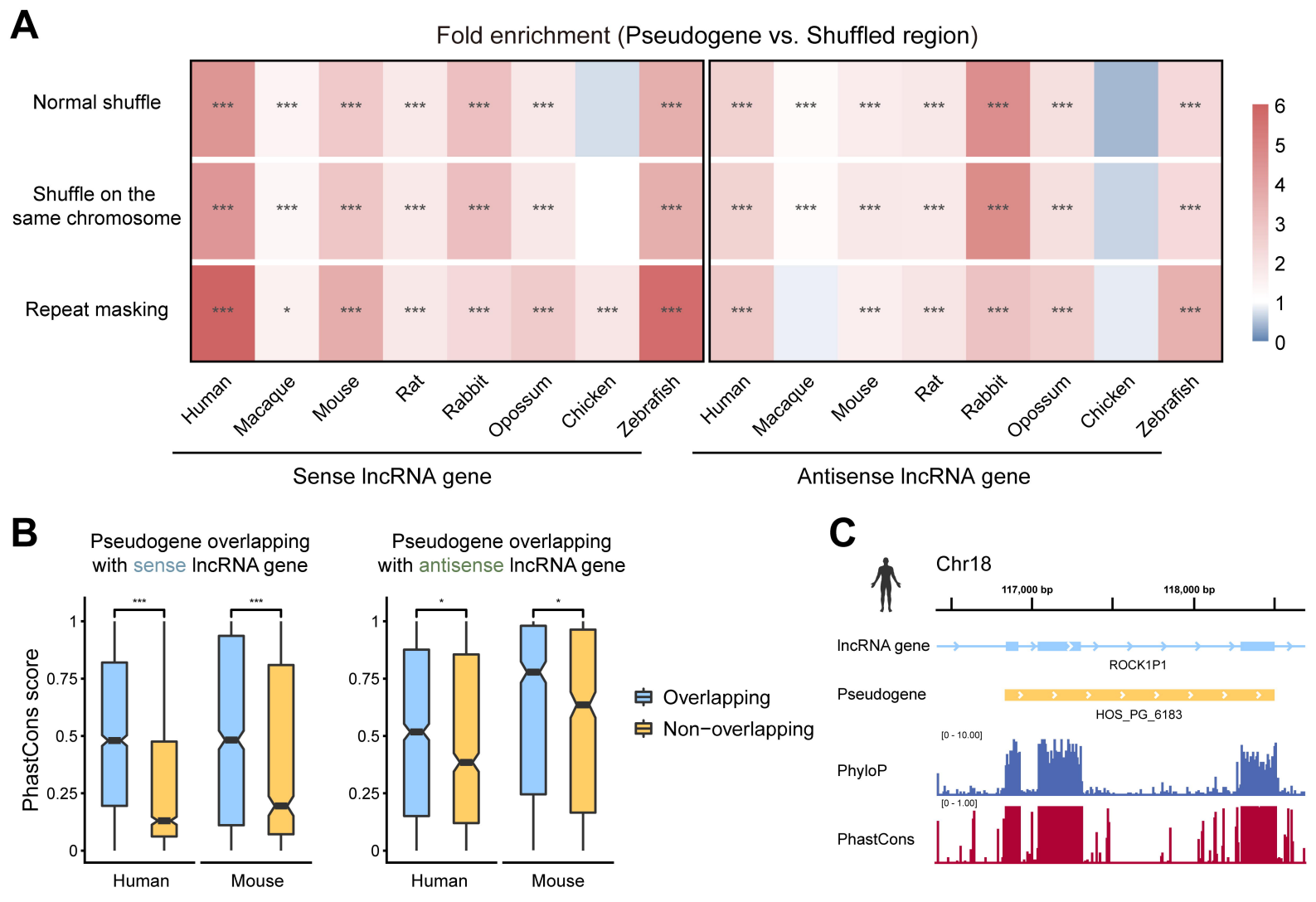
1444 **A**. Enrichment of evolutionary ages. **B**. Exonic sequence conservation (PhastCons
1445 score). **C**. Transcript lengths.
1446
1447 **Figure S52 Functionality characterization of PAS lncRNA genes at a series of**
1448 **expression level thresholds in mouse**
1449 **A**. Enrichment of evolutionary ages. **B**. Exonic sequence conservation (PhastCons
1450 score). **C**. Transcript lengths.
1451
1452 **Figure S53 Functionality characterization of PAA lncRNA genes at a series of**
1453 **expression level thresholds in mouse**
1454 **A**. Enrichment of evolutionary ages. **B**. Exonic sequence conservation (PhastCons
1455 score). **C**. Transcript lengths.
1456
1457 **Table S1 Pseudogene-associated lncRNA genes in eight vertebrates**
1458
1459 **Table S2 Functionality characterization of pseudogene-associated lncRNA**
1460 **genes in human and mouse**
1461
1462 **Table S3 Sample information of the public single-cell/nucleus RNA sequencing**
1463 **data**
1464
1465 **Table S4 Abbreviations, marker genes, and number for each cell type in each**
1466 **tissue**
1467
1468 **Table S5 Aging-related differential expression of pseudogene-associated**
1469 **lncRNA genes in human**
1470
1471 **Table S6 Aging-related differential expression of pseudogene-associated**
1472 **lncRNA genes in mouse**

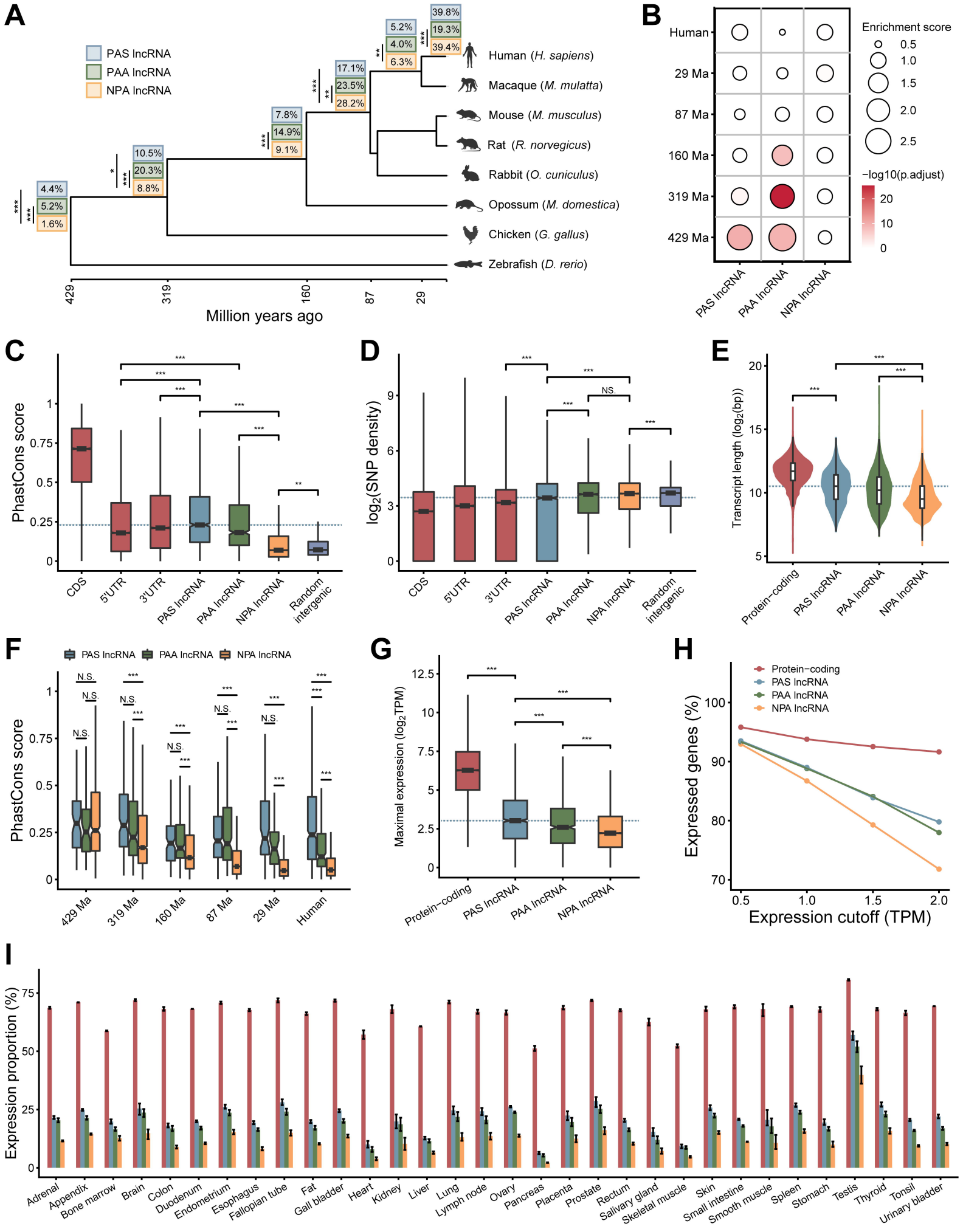
1473

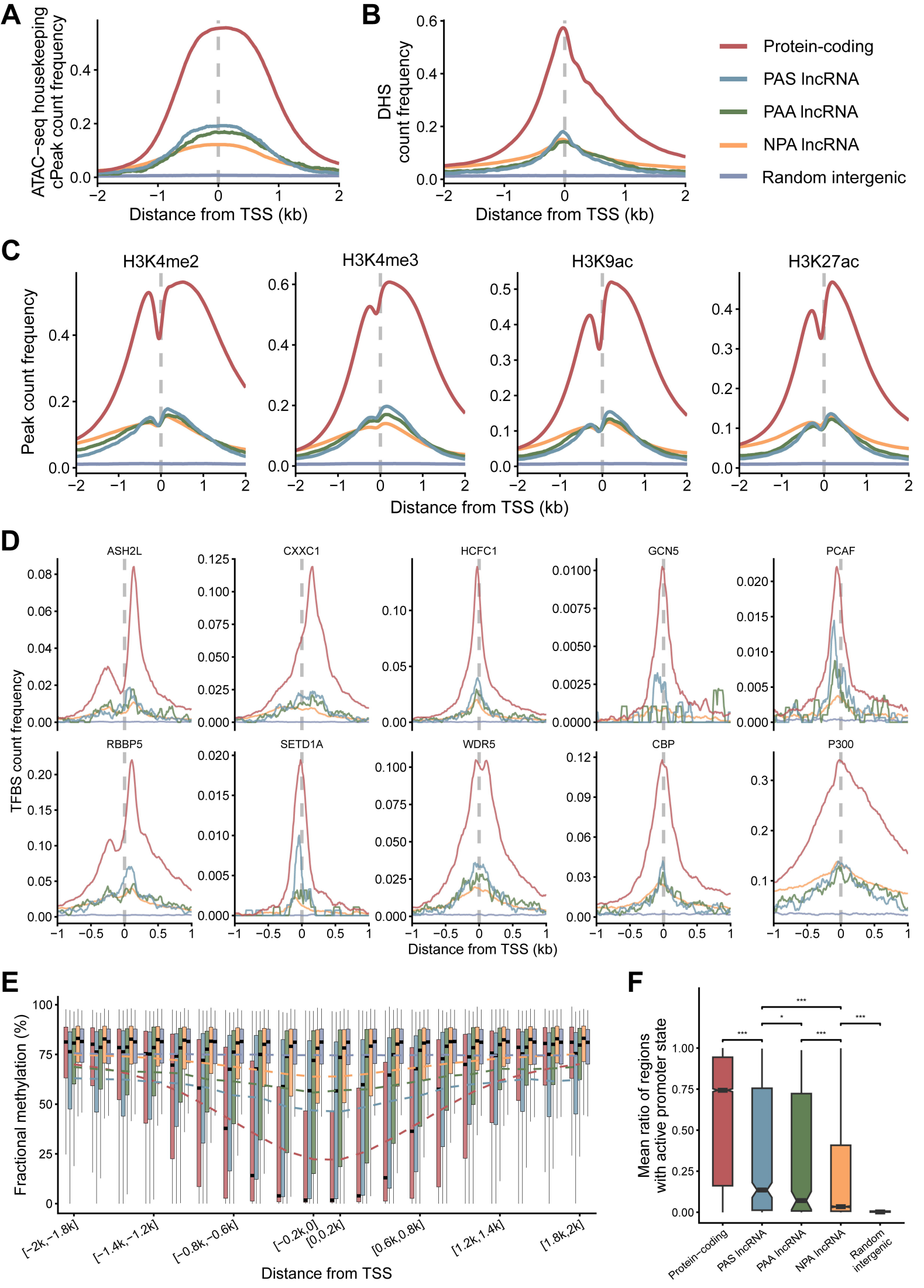
1474 **Table S7 Primers for quantitative reverse transcription PCR**

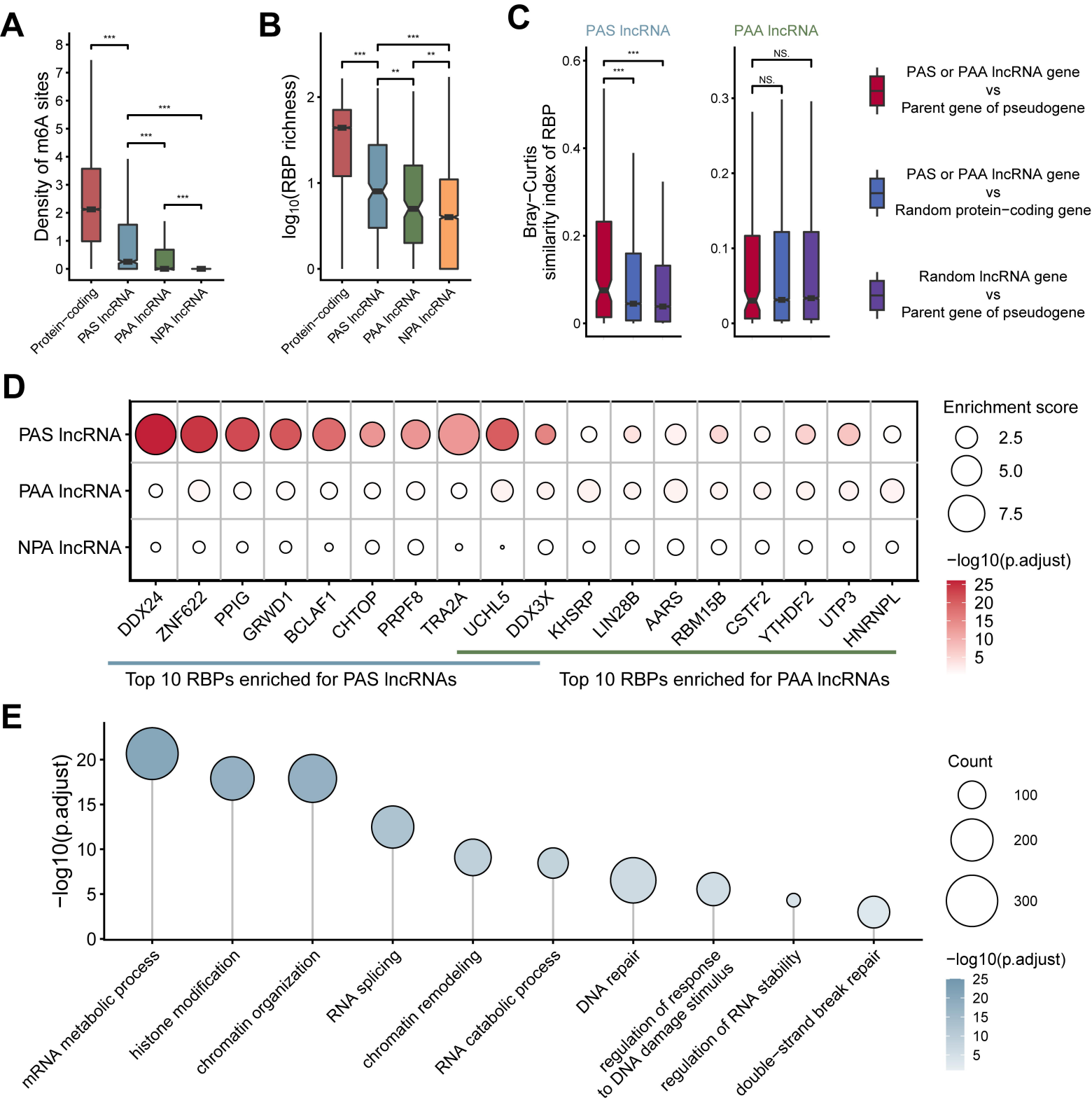
1475

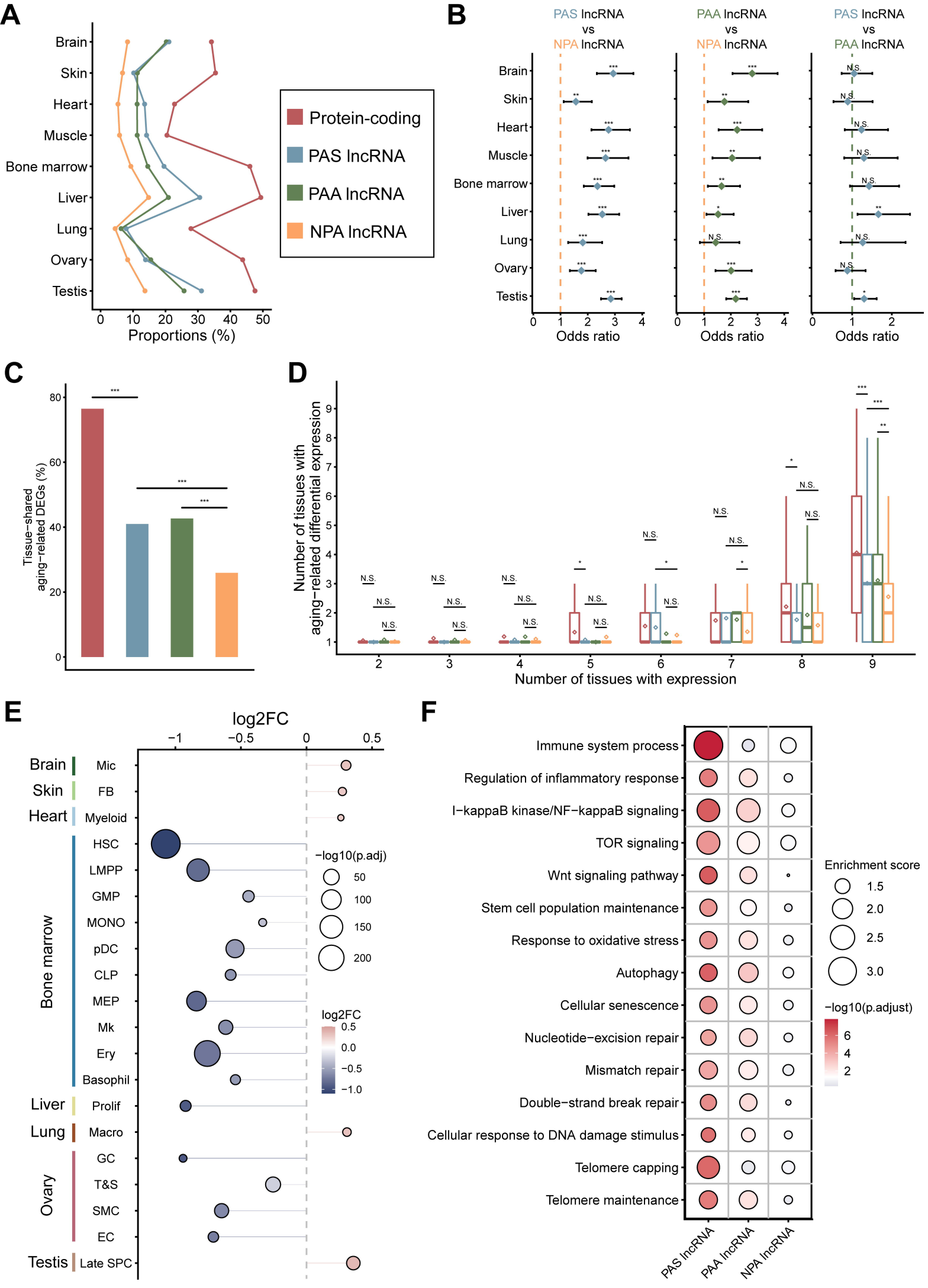

A
Species
Mouse
Chicken
Zebrafish
Sexes
Male
Female
Tissues
Brain & Cerebellum
Heart
Gut
Testis
Ovary
Illumina
PacBio
B
Zebrafish
Chicken
Mouse
Transcript length (log10(bp))
PacBio
ENSEMBL
C
Transcript exon number (log10(count))
D
Human
Macaque
Mouse
Rat
Rabbit
Opossum
Chicken
Zebrafish
Novel
ENSEMBL
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

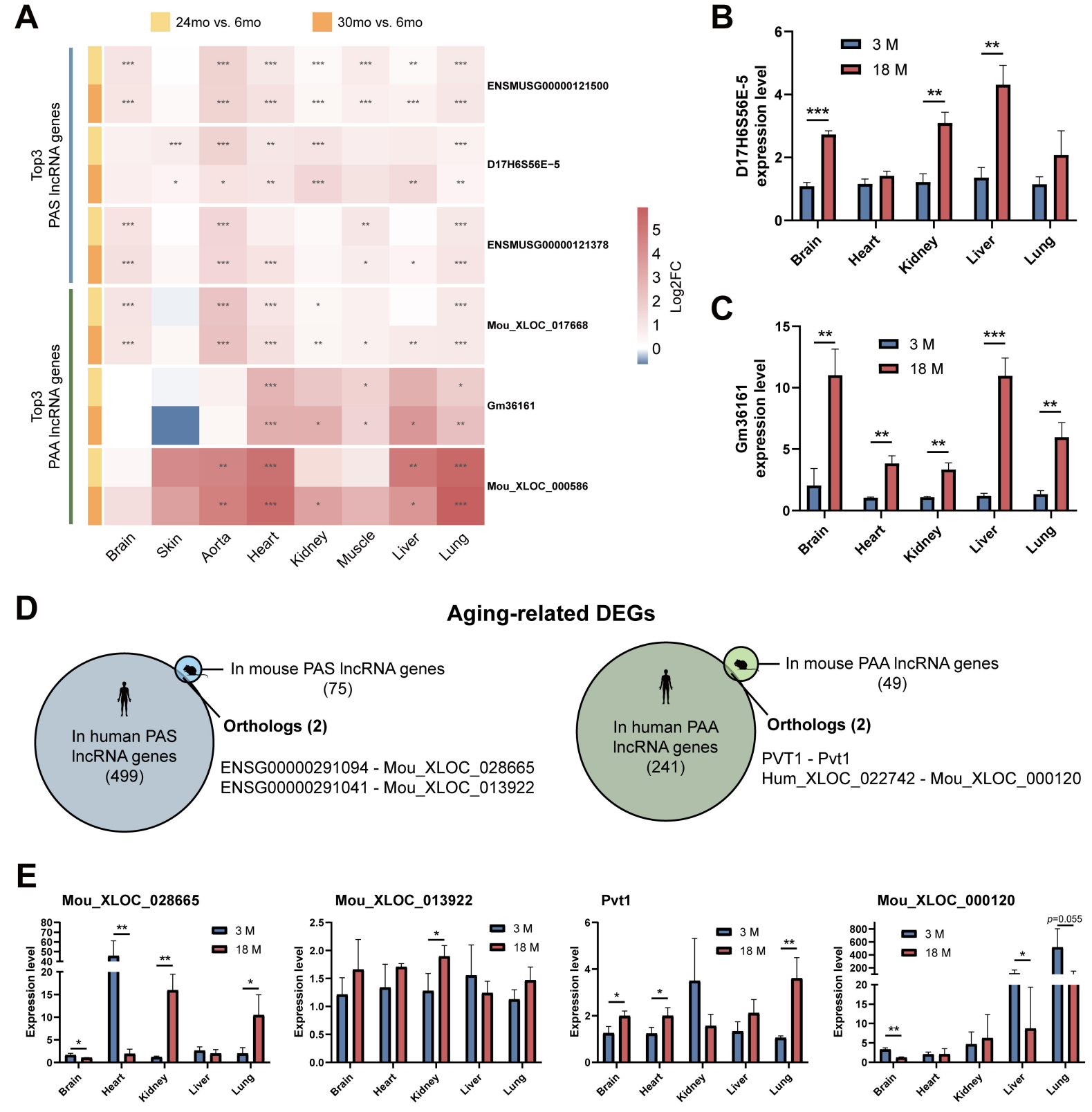

A
Fold enrichment (Pseudogene vs. Shuffled region)
Normal shuffle
Shuffle on the same chromosome
Repeat masking
Human
Macaque
Mouse
Rat
Rabbit
Opossum
Chicken
Zebrafish
Sense lncRNA gene
Antisense lncRNA gene
6
5
4
3
2
1
0
B
Pseudogene overlapping with sense lncRNA gene
Pseudogene overlapping with antisense 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
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
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
Brain
Mic
Skin
FB
Heart
Myeloid
Bone marrow
HSC
LMPP
GMP
MONO
pDC
CLP
MEP
Mk
Ery
Basophil
Liver
Prolif
Lung
Macro
Ovary
GC
T&S
SMC
EC
Testis
Late SPC
−log10(p.adj)
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)
Orthologs (2)
In human PAS lncRNA genes (499)
ENSG00000291094 - Mou_XLOC_028665
ENSG00000291041 - Mou_XLOC_013922
In mouse PAA lncRNA genes (49)
Orthologs (2)
In human PAA lncRNA genes (241)
PVT1 - Pvt1
Hum_XLOC_022742 - Mou_XLOC_000120
E
Mou_XLOC_028665
Mou_XLOC_013922
Pvt1
Mou_XLOC_000120
Expression level
p=0.055