Gene fate spectrum as a reflection of local genomic properties

Yuichiro Hara[1*] and Shigehiro Kuraku[2,3,4]

[1]Research Center for Genome & Medical Sciences, Tokyo Metropolitan Institute of Medical Science, Japan

[2]Molecular Life History Laboratory, Department of Genomics and Evolutionary Biology, National Institute of Genetics, Japan

[3]Department of Genetics, Sokendai (Graduate University for Advanced Studies), Japan

[4]Laboratory for Phyloinformatics Team, RIKEN Center for Biosystems Dynamics Research (BDR), Japan

[*]Correspondence: hara-yi@igakuken.or.jp

Abstract

Functionally indispensable genes are likely to be retained and otherwise to be lost during evolution. This evolutionary fate of a gene can also be affected by neutral factors, including the mutability of genomic positions, but such features have not been examined well. To uncover the genomic features associated with gene loss, we investigated the characteristics of genomic regions where genes have been independently lost in multiple lineages. With a comprehensive scan of gene phylogenies of vertebrates with a careful inspection of evolutionary gene losses, we identified 1,081 human genes whose orthologs were lost in multiple mammalian lineages: designated 'elusive genes.' These elusive genes were located in genomic regions with rapid nucleotide substitution, high GC content, and high gene density. A comparison of the orthologous regions of such elusive genes across vertebrates revealed that these features had been established before the radiation of the extant vertebrates more than 500 million years ago. The association of human elusive genes with transcriptomic and epigenomic characteristics illuminated that the genomic regions containing such genes were subject to repressive transcriptional regulation. Thus, the heterogeneous genomic features driving gene fates toward loss have been in place since the ancestral vertebrates and may sometimes have relaxed the functional indispensability of such genes.

Introduction

In the course of evolution, genomes continue to retain most genes with occasional duplications, while losing some genes. This retention and loss can be interpreted as gene fate; genes are stably retained in the genome, but some factors may cause them to transition to a state where deletion occurs. Accordingly, identification of the factors allowing gene loss may facilitate our understanding of gene fate. Gene retention or loss has generally been considered to depend largely on the functional importance of the particular gene from the perspective of molecular evolutionary biology (Albalat and Cañestro, 2016; Bartha et al., 2018; Blanc et al., 2012; Liu et al., 2015; Olson, 1999; Sharma et al., 2018; Shen et al., 2018). Genes with indispensable functions have usually been retained with highly conserved sequences in genomes, through rapid elimination of alleles that impair gene functions (Hirsh and Fraser, 2001; Krylov et al., 2003; Miyata et al., 1980; Pál et al., 2006). However, genes with less important functions are likely to accept more mutations and structural variations, which can degrade the original functions, leading to gene loss through pseudogenization or genomic deletion (Jordan et al., 2002; Yang et al., 2003). To date, gene loss has been imputed to the relaxation of functional constraints of individual genes. Gene loss has further been revealed to drive phenotypic adaptation in various organisms (Albalat and Cañestro, 2016; Olson, 1999), as well as in a gene knockout collection of yeasts in culture (Giaever and Nislow, 2014; Maclean et al., 2017).

To uncover the association between fates and functional importance of the genes, molecular evolutionary analyses have been conducted at various scales, from gene-by-gene to genome-wide. A number of studies have revealed that the genes with reduced non-synonymous substitution rates (or K_A values) and ratios of non-

synonymous to synonymous substitution rates (K_A/K_S ratios) are less likely to be lost (Jordan et al., 2002; Yang et al., 2003). A genome-wide comparison of duplicated genes in yeast revealed larger K_A values for those lost in multiple lineages than those retained by all the species investigated (Byrne and Wolfe, 2007). Other comprehensive studies of gene loss across metazoans and teleosts revealed that the genes expressed in the central nervous system are less prone to loss (Fernández and Gabaldón, 2020; Roux et al., 2017). These observations again suggest that gene fate depends on the functional constraints of a particular gene.

Besides functional constraints, several studies have identified the genes lost independently in multiple lineages, revealing that the genomic regions containing these genes 'prefer' particular characteristics associated with structural instability (Cortez et al., 2014; Hughes et al., 2012; Lewin et al., 2021; Maeso et al., 2016). In mammals, tandemly arrayed homeobox genes derived from the Crx gene family were lost in multiple species (Lewin et al., 2021; Maeso et al., 2016). The findings suggest that genomic features containing tandem duplications facilitate unequal crossing over, leading to frequent gene loss. Mammalian chromosome Y, which contains abundant repetitive elements and continues to reduce in size, has lost a considerable number of genes (Cortez et al., 2014; Hughes et al., 2012). Genes in such particular genomic regions may be prone to loss in a more neutral manner than the relaxation of functional importance or via functional adaptations. Accordingly, these studies focusing on the particular genomic regions led us to search for the common features in genomes that potentially facilitate gene loss. Genome-wide scans have revealed heterogeneous distributions of a variety of sequence and structural features so far, for example, base composition (Bernardi and Bernardi, 1986; Cohen et al., 2005; Katzman et al., 2011),

the frequency of repetitive elements (Korenberg and Rykowski, 1988; Medstrand et al., 2002), and DNA-damage sensitivity induced by replication inhibitors (Debatisse et al., 2012; Helmrich et al., 2006). However, the extent to which these characteristics are associated with gene fates has not been understood well at a genome-wide level.

The accumulation of near-complete genome assemblies for various organisms facilitates comprehensive taxon-wide analysis of gene loss (Fernández and Gabaldón, 2020; Guijarro-Clarke et al., 2020; Rice and McLysaght, 2017). Along with this motivation, we recently performed a comprehensive analysis on the fate of paralogs generated via the two-round whole genome duplications in early vertebrates (Hara et al., 2018a). The results revealed that the genes retained by reptiles but lost in mammals and Aves rapidly accumulated not only non-synonymous but also synonymous substitutions in comparison with the counterparts retained by almost all the vertebrates examined, indicating that those genes prone to loss harbor rapid mutation rates. Furthermore, these loss-prone genes were located in genomic regions with high GC-contents, high gene densities, and high repetitive element frequencies. These findings suggest that the fates of those genes are influenced not only by functional constraints but also by intrinsic genomic characteristics. Because the findings were restricted to a set of particular genes, they prompted us to examine whether this trend is associated with gene fates on a genome-wide scale.

In this study, we inferred molecular phylogenies of vertebrate orthologs to systematically search for the genes harboring different fates in the human genome. We referred to the loss-prone genes as ‘elusive’ genes that were retained by modern humans but were lost independently in multiple mammalian lineages. As a comparison of the elusive genes, we retrieved the ‘non-elusive’ genes that were retained by almost all of

the mammalian species examined. We conducted a careful search for gene loss to reduce the false discovery rate, which is usually caused by incomplete sequence information (Botero-Castro et al., 2017; Deutekom et al., 2019). By comparing the genomic regions containing these genes, we uncovered genomic characteristics relevant to gene loss. We associated the elusive genes with a variety of findings from deep sequencing analyses of the human genome including transcriptomics, epigenomics, and genetic variations. These data assisted us to understand how intrinsic features of genomes—presumably unrelated to gene function, may affect gene fate, leading to loss by relaxing the functional importance of 'elusive' genes.

Results

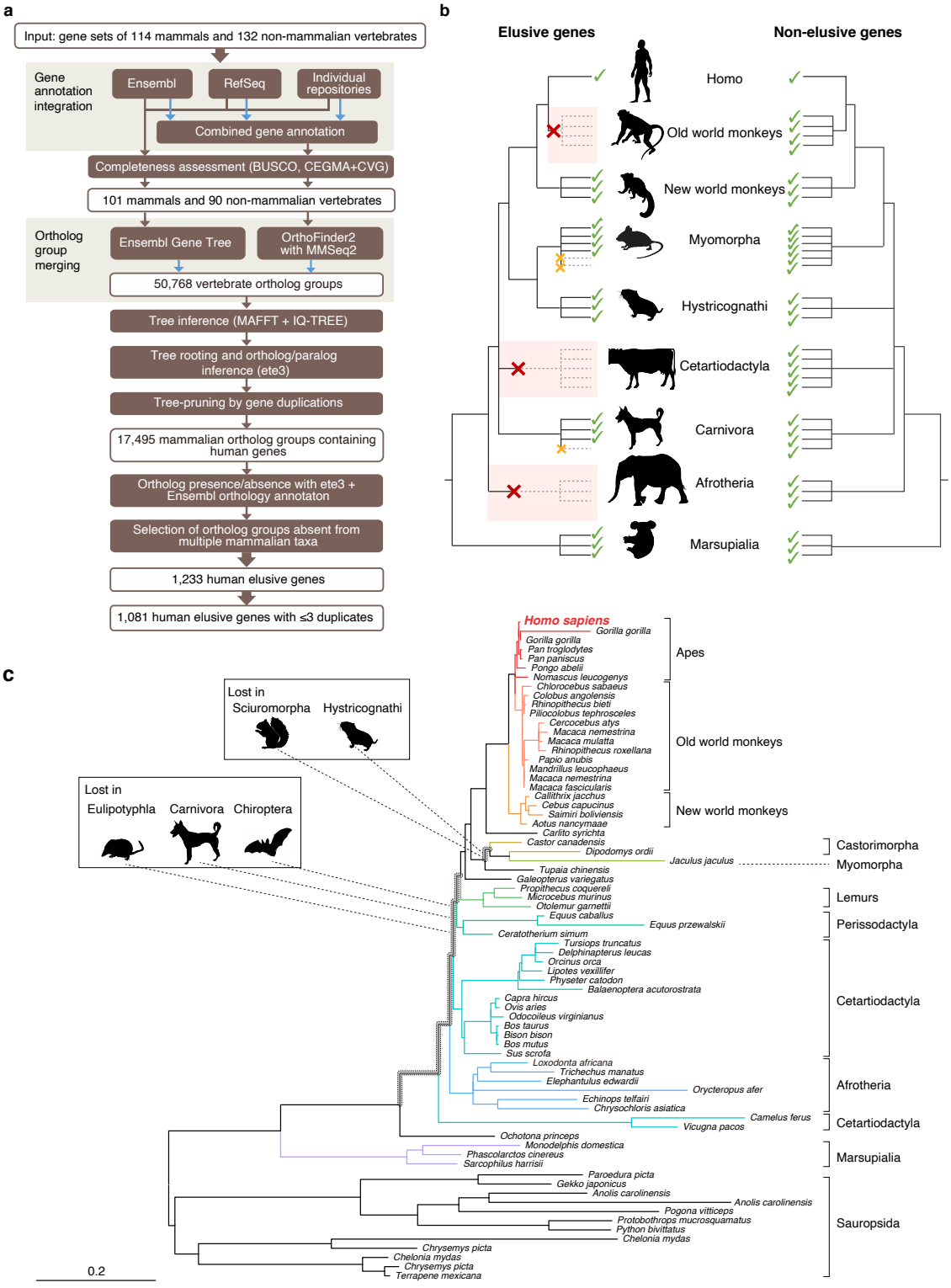
Identification of human 'elusive' genes

We defined an 'elusive' gene as a human protein-coding gene that existed in the common mammalian ancestors but was lost independently in multiple mammalian lineages (Figure 1; see Methods for details). In our analysis, we searched for such genes by reconstructing phylogenetic trees of vertebrate orthologs and detecting gene loss events within the individual trees. To search for elusive genes, we paid close attention to distinguishing true evolutionary gene loss from falsely inferred gene loss caused by insufficient genome assembly, gene prediction, and orthologous clustering (Botero-Castro et al., 2017; Deutekom et al., 2019), as described below.

We first produced highly complete orthologous groups comprised of nearly complete gene sets. We merged multiple gene annotations of a single species followed by assessments of the completeness of the gene sets (Figure 1a). Using these gene sets, we then created two sets of ortholog groups with different methods and merged them

into a single set (Figure 1a). In searching for gene loss events, we restricted our study to those that occurred in the common ancestors of particular 'higher' taxa. This procedure relieved false identifications of gene loss in a species or an ancestor of a lower taxonomic hierarchy caused by incomplete genomic information (Figure 1b).

We integrated gene annotations from Ensembl, RefSeq, and the sequence repositories of individual genome sequencing projects to produce gene annotations for 114 mammalian and 132 non-mammalian vertebrates. From these, we selected the annotations of 101 and 90 species, respectively, that exhibited high completeness in the BUSCO assessment (Simão et al., 2015) (Supplementary Table S1). Using these gene sets, ortholog clustering was conducted by OrthoFinder, and these ortholog groups were integrated into the ones provided by the Ensembl Gene Tree. This integration resulted in 50,768 vertebrate ortholog groups. Phylogenetic tree inference of the integrated ortholog groups and pruning of the individual trees based on gene duplications resulted in 17,495 mammalian ortholog groups that contained human genes. For the individual mammalian ortholog groups, we searched for family or 'higher' taxonomic groups (listed in Supplementary Table S1) in which the gene was absent in all the species examined (Figure 1b). We interpreted this gene absence as an evolutionary loss that occurred in the common ancestor of the taxon. Finally, we extracted the ortholog groups that were retained by humans but were lost independently in the common ancestors of at least two taxa (Figure 1c). Hereafter we call the human genes belonging to these ortholog groups 'elusive genes.' To compare these, we also selected the ortholog groups that contained all of the mammals examined including single-copy human genes. We called these 'non-elusive genes.' This comprehensive scan of gene phylogenies resulted in 1,081 elusive and 8,050 non-elusive genes (Supplementary Table S2).



155

Figure 1. Detection of 'elusive' genes

(**a**) Pipeline of ortholog group clustering and gene loss detection. (**b**) Definition of an elusive gene schematized with ortholog presence/absence pattern referring to a 'higher' taxonomic hierarchy. (**c**) A representative phylogeny of the elusive gene encoding Chitinase 3-like 2 (CHI3L2). Taxa shown in the tree were used to investigate the presence or absence of orthologs. The Sciuromorpha, Hystricognathi, Eulipotyphla, Carnivora, and Chiroptera are absent from the tree, indicating that the CHI3L2 orthologs were lost somewhere along the branches framed in gray in the tree. In addition, the orthologs of many members of the Myomorpha were not found, suggesting that gene loss occurred in this lineage.

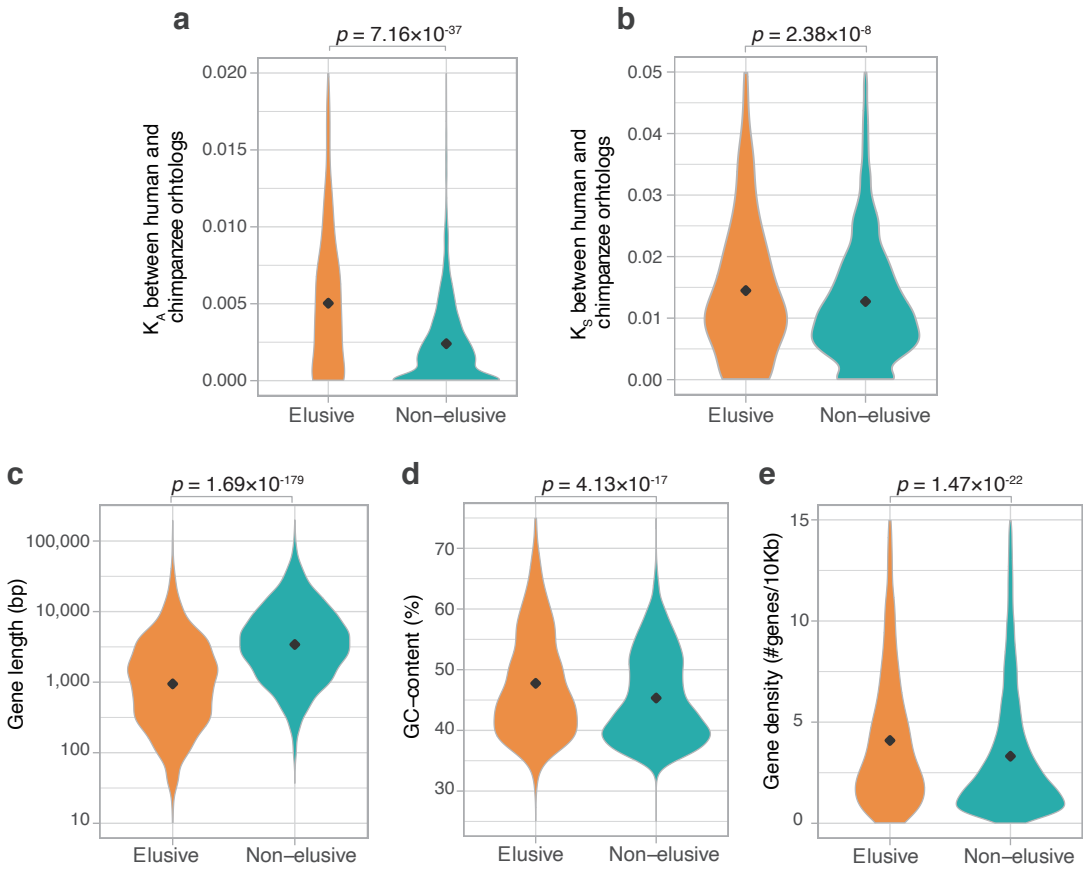
Genomic signatures of the human elusive genes

The loss-prone nature of the elusive genes suggests a relaxation of their functional constraints. To uncover the molecular evolutionary characteristics associated with each elusive gene, we computed synonymous and non-synonymous substitution rates, namely K_S and K_A values, respectively, between human and chimpanzee and mouse orthologs for the elusive and non-elusive genes. The results showed larger K_A values in the ortholog pairs of the elusive genes than in those of the non-elusive genes (Figure 2a; Figure 2–figure supplement 1). This indicates rapid accumulation of amino acid substitutions in the elusive genes, potentially accompanied by the relaxation of functional constraints. Our analysis further illuminated larger K_S values for the elusive genes than in the non-elusive genes (Figure 2b; Figure 2–figure supplement 1). Importantly, the abundance of synonymous substitutions, which do not affect changes in amino acid residues, indicates that the elusive genes are also susceptible to genomic characteristics independent of selective constraints on gene functions.

To further scrutinize the characteristics reflecting the genomic environment rather than gene function, we analyzed genomic characteristics that may distinguish the elusive from non-elusive genes. A comparison between these two categories revealed shorter gene-body lengths and higher GC contents of elusive rather than non-elusive genes (Figure 2c,d). Furthermore, a scan of intergenomic gene distribution revealed that the elusive genes were located in the genomic regions with high gene density compared with the non-elusive genes (Figure 2e). Our findings indicate that such elusive genes have distinct characteristics in the human genome. These genomic characteristics, as well as high nucleotide substitution rates, were consistent with the findings in our

189 genome analyses using the amniote and elasmobranch genomes (Hara et al., 2018b,

190 2018a).



191

192 **Figure 2. Genomic and evolutionary characteristics of elusive genes**

193 Distributions of non-synonymous and synonymous substitution rates, namely K_A (**a**)

194 and K_S (**b**) values, respectively, between the human-chimpanzee orthologs of the elusive

195 and non-elusive genes. Distribution of gene length (**c**) and GC content (**d**) of the human

196 elusive and non-elusive genes. (**e**) Distribution of gene density in the genomic regions

197 where the human elusive and non-elusive genes are located.

198

Tracing elusiveness back along the vertebrate evolutionary tree

The origins of the human elusive genes can be traced back along the evolutionary tree, at least to the mammalian common ancestor. To investigate possible antiquities of the genomic properties associated with elusive genes, we investigated their orthologs in non-mammalian vertebrate genomes. By scrutinizing the ortholog groups that were used for elusive gene identification, we identified 982 human elusive gene orthologs for chimpanzee, 540 for mouse, 380 for chicken, 415 for central bearded dragon, 416 for clawed frog, 415 for coelacanth, 431 for spotted gar, and 390 for bamboo shark. These four non-mammalian vertebrates retained orthologs of fewer than half of the elusive genes, but most of the non-elusive ones (Figure 3–figure supplement 1a). In the coelacanth, gar, and shark, the orthologs of the elusive genes were less frequently retained by all the species than those of the non-elusive ones (Figure 3–figure supplement 1b). This suggests that the origins of the loss-prone propensity of the elusive genes potentially date back to long before the emergence of the Mammalia.

We further examined the genomic characteristics harbored by the human elusive genes in the vertebrate orthologs. In all the species examined, the orthologs of the elusive genes exhibited high GC content and compact gene bodies. Additionally, in most of these species, the orthologs of elusive genes were located in genomic regions with high gene density compared with orthologs of the non-elusive genes (Figure 3; Figure 3–figure supplement 2). In addition, we computed K_S and K_A values between the orthologs of the vertebrate species and their close relatives for elusive and non-elusive genes. In any of the species pairs, the orthologs of the elusive genes were found to harbor higher K_S values than those of the non-elusive gene orthologs, while the orthologs of the elusive genes exhibited higher K_A values in mammals and lizards

223 (Figure 3; Figure 2–figure supplement 1). These observations indicate that these
224 genomic characteristics probably originated before the emergence of gnathostomes, a
225 monophyletic group of chondrichthyan and bony vertebrates, and have been retained for
226 at least 500 million years.
227

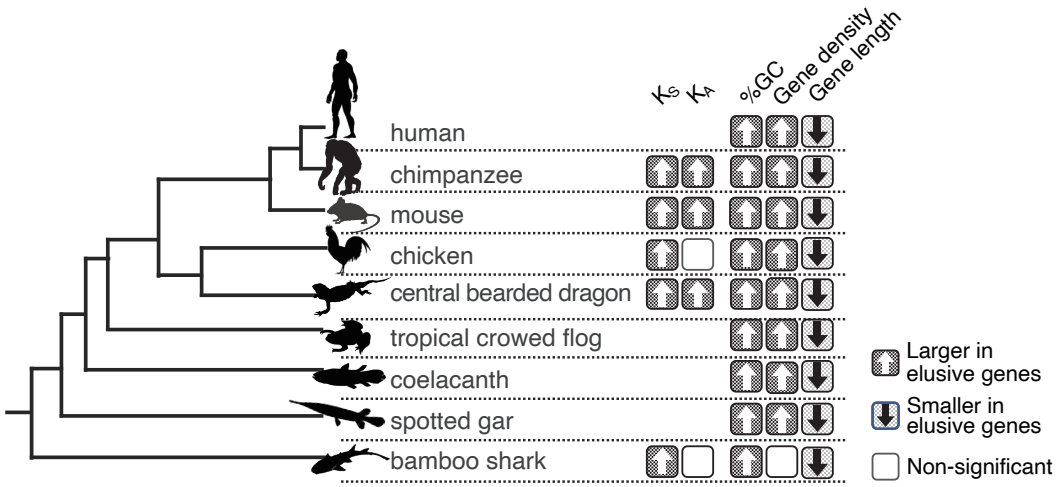


Figure 3. Longstanding characteristics of elusive genes

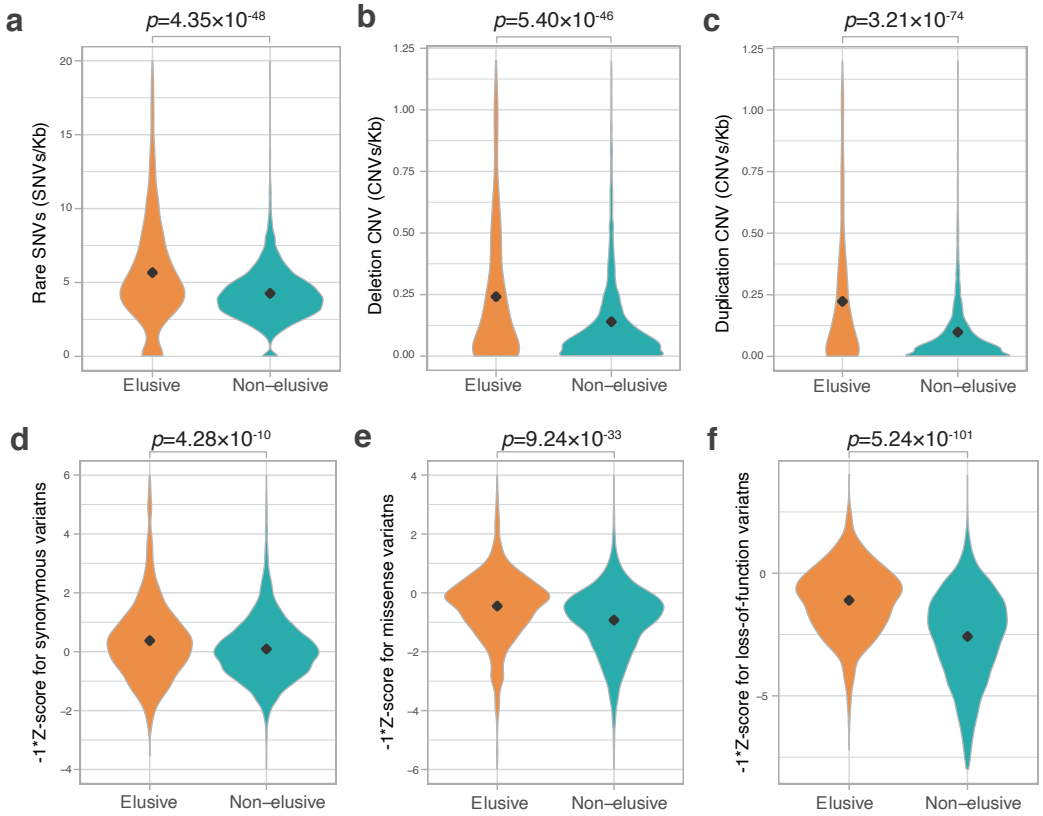
Retention of the genomic and evolutionary characteristics of the human elusive genes across vertebrates. The individual round squares with arrowheads indicate significant increases or decreases of the distribution of particular characteristics in the orthologs of the human elusive genes and their flanking regions, compared with those of the non-elusive genes in these selected vertebrate genomes. For the chimpanzee and mouse genomes, K_A and K_S values were computed between the human elusive genes and the orthologs of these mammals. For the non-mammalian species, these values were computed with ortholog pairs for the elusive/non-elusive genes between the corresponding species and their closely related species: turkey for chicken, green anole for central bearded dragon, and whale shark for bamboo shark. Distributions of these metrics for non-human species are shown in Supplementary Figures S1 and S3.

Abundant polymorphism in elusive genes

The observation of large K_S and K_A values in the elusive genes prompted us to examine the extent to which these genes have accommodated genetic variations in modern humans. Large-scale human genome resequencing projects have identified a huge number of genetic variations, from rare to common, and from single nucleotide variants (SNVs) to chromosome-scale structural variants, facilitating tackling this issue. We retrieved copy number variants (CNVs) and rare SNVs in the human genome from the Database of Genomic Variants, release 2016-08-31 (MacDonald et al., 2014) and dbSNP release 147 (Sherry et al., 2001), respectively, and computed their densities in the individual genic regions. We found that the genic regions of the human elusive genes contained abundant rare SNVs, as well as deletion and duplication CNVs, compared with those of the non-elusive genes (Figure 4a–c). This result suggests that genomic regions containing the elusive genes are not only prone to loss but also to duplication.

To evaluate the functional consequences of abundant genetic variants in the elusive genes, we investigated genetic variations stored in the gnomAD v. 2.1 database, a repository containing >120,000 exome and >15,000 whole genome sequences of human individuals (Karczewski et al., 2021). This database classifies SNVs in coding regions into three categories—synonymous, missense, and loss-of-function—and the loss-of-function category contains nonsense mutations, frameshift mutations, and mutations in splicing junctions. The gnomAD site computes a *Z*-score, an index representing the abundance of SNVs for individual genes; positive and negative values denote fewer or more mutations in a coding region than expected, respectively (Figure 4d–f). Accordingly, the *Z*-score for nonsense mutations and loss-of-function mutations of the

267 individual genes indicates the degree of natural selection: larger values demonstrate
268 genes subjected to purifying selection, while smaller ones suggest functional relaxation.
269 We found lower *Z*-scores of missense and loss-of-function mutations (higher opposite
270 numbers of *Z*-scores in Figure 4e, f) in the human elusive genes than in the non-elusive
271 genes, suggesting that the elusive genes are more functionally dispensable and
272 potentially resistant to harmful mutations. Additionally, the *Z*-scores of synonymous
273 mutations of the human elusive genes were higher than those of the non-elusive genes
274 (Figure 4d). This confirms the high mutability of genomic regions containing elusive
275 genes, as observed in the K_S values.



276

277 **Figure 4. Genetic variations of the elusive and non-elusive genes within human**

278 **populations**

279 Comparison of the density of rare SNVs (**a**), deletion CNVs (**b**), duplication CNVs (**c**),

280 and Z-scores of synonymous (**d**), missense (**e**), and loss-of-function variants (**f**). We

281 used opposite numbers of the *Z*-scores in **d–f** so that the elusive genes have higher

282 values than non-elusive genes as in Figures 2a ,b, d, e and 3a–c.

283

Transcriptomic natures of elusive genes

To further investigate how the human elusive genes have decreased functional essentiality, we examined their expression profiles. For this purpose, we compared gene expression profiles of the 54 adult tissues from the GTEx database v. 8 (GTEx Consortium, 2020) between the elusive and non-elusive genes. For individual genes, we computed the maximum Transcription Per Million (TPM) values among these tissues as the expression quantity level. For expression diversities, we employed Shannon's diversity index H', which is often utilized as an index of species diversity in the ecological literature, based on the proportion of TPM values across the 54 tissues.

As shown in the density scatter plots of the individual genes displaying these two indicators in Figure 5, most of the non-elusive genes possessed large maximum TPM and H' values. Thus, most non-elusive genes are ubiquitously expressed at certain levels. By contrast, the density plot of the elusive genes displayed an additional high-density spot with small TPM and H' values, indicating that the genes in this spot were not expressed, at least in adult tissues. The plot also showed another broad dense area of small H' values, which contained the genes expressed in a single or a few tissues. A similar analysis was performed with the fetal single cell RNA-seq data (Cao et al., 2020), revealing that the averaged expression profiles of the elusive and non-elusive genes for the 172 cell types were concordant with those of the adult tissues (Figure 5). Our findings demonstrate that some elusive genes harbor low-level and spatially-restricted expression profiles, which are rarely observed in the non-elusive genes.

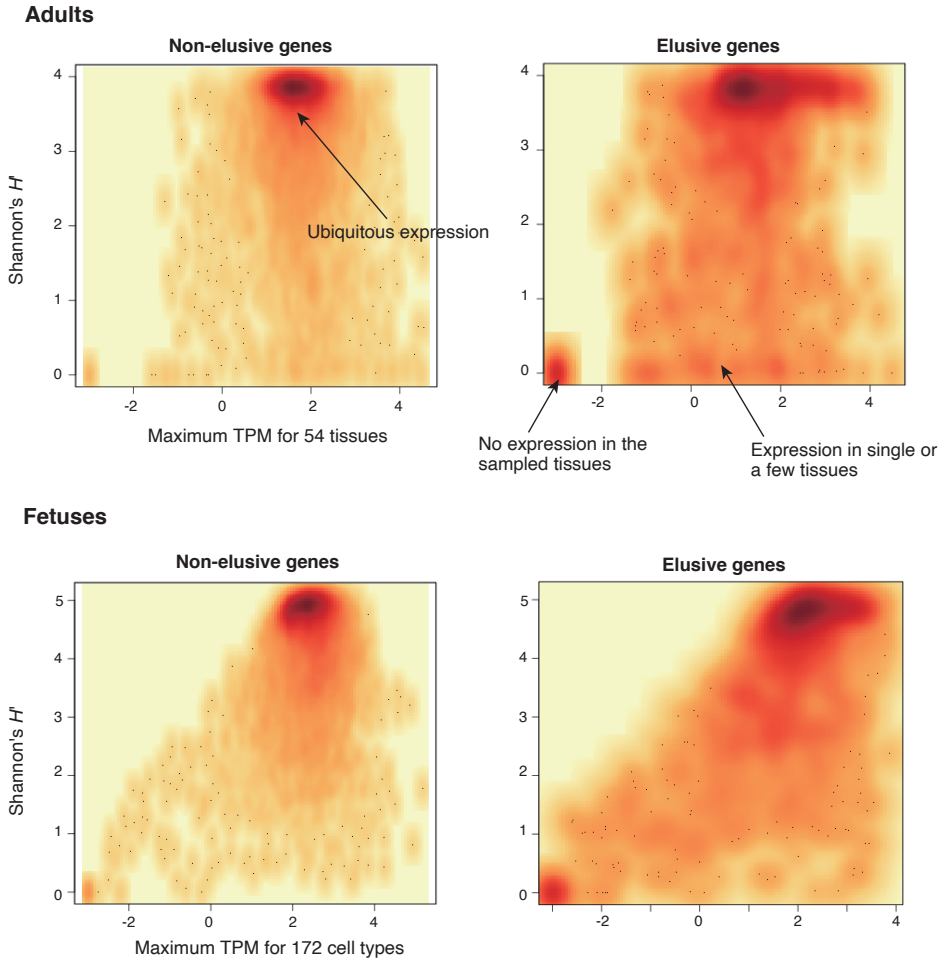


Figure 5. Expression profiles of elusive and non-elusive genes

The figure shows density scatter plots of the expression quantity and divergence of elusive and non-elusive genes. The median TPM values of the individual adult tissues across populations were retrieved from the GTEx database (GTEx Consortium, 2020), and normalized TPM values of the fetal cell types were retrieved from the Descartes database (Cao et al., 2020). For the individual genes, maximum TPM and Shannon's *H'* values were computed using these processed TPM values.

Epigenetic nature of elusive genes

Our finding of the low-level and spatially-restricted expression patterns of elusive genes prompted us to explore epigenetic properties involved in this transcriptional regulation. Therefore, we retrieved epigenetic data on a variety of human cell lines from a few regulatory genome databases including ENCODE, a repository that stores the comprehensive annotations of functional elements in the human genome (ENCODE Project Consortium, 2012). Using this information, we characterized the epigenetic features of the genomic regions containing elusive genes (Figure 6).

We compared peak densities based on the Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq), an indicator of accessible chromatin regions in the genome, in the gene bodies and flanking regions between the elusive and non-elusive genes. In seven cell lines out of eight examined (nine experiments of ten), the results showed fewer ATAC-seq peaks in the genomic regions including the elusive genes than in those including non-elusive genes, indicating that the elusive genes are likely to reside in inaccessible genomic regions (Figure 6a; Figure 6–figure supplement 1). We also searched for Topologically Associating Domains (TADs), genomic elements with frequent physical self-interaction potentially acting as promoter-enhancer contacts (Rao et al., 2014) that included either the elusive or non-elusive genes. The result showed that a higher fraction of the elusive genes resided outside of the TADs than the non-elusive genes for all the eleven cell lines investigated (Figure 6b; Figure 6–figure supplement 2). Furthermore, the elusive genes were located in shorter TADs. These observations suggest that the elusive genes are unlikely to be regulated by distant regulatory elements (Figure 6b).

Our investigations extended to the association of the elusive genes with further global regulation of genomic structures. We compared the percentage normalized signal of Repli-seq (Hansen et al., 2010), a high throughput sequencing for quantifying DNA replication time as a function of genomic position, between the elusive and non-elusive genes. The results showed that elusive genes were prone to late replication in all of the 15 cell lines examined (Figure 6c; Figure 6–figure supplement 1). Late-replicating regions are frequently located at the nuclear periphery and often interact with the nuclear lamina. Therefore, we examined the nuclear position of the genomic regions including the elusive genes by referring to the Lamina Associating Domains (LADs) that were identified by the ChIP-seq reads for Lamin B1 (van Schaik et al., 2020; Zheng et al., 2018). Compared with the non-elusive genes, the elusive genes were found to be enriched in LADs for all of the four cell lines examined (Figure 6d; Figure 6–figure supplement 3), consistent with their late replication timings (van Steensel and Belmont, 2017).

We further investigated the association of the restricted expressions of the elusive genes with epigenetic features. From 988 elusive genes whose expressions were quantified in the GTEx database, we classified the elusive genes into two groups based on the expression diversities: that is, 173 elusive genes with Shannon's diversity index $H' > 1$ were ubiquitously expressed, and 815 of those with $H' \leq 1$ were expressed in only a few or none of the tissues examined (Figure 5). Importantly, all of the four epigenetic features of the elusive genes with $H' \leq 1$ were more pronounced than those with $H' > 1$: sparse ATAC-seq peaks, short TADs, late replication timings, and significant overlaps with LADs (Figure 6; Figure 6–figure supplement 4). This

observation suggests that low-level and spatially-restricted expressions of the elusive genes are associated with intrinsic epigenetic features of these genomic regions.

High GC contents in genomic regions potentially hinder identifying an epigenetic feature by short read sequencing because of underrepresentation of sequence reads by amplification-based sequencing libraries. This bias might lead to sparse distributions of the ATAC-seq peaks and Hi-C contacts in the genomic regions that contain the elusive genes. However, only 8.09% and 10.8% of the elusive genes with $H' \leq 1$ and $H' > 1$ were located in extremely high GC-content regions (>60%), respectively, with no significant difference ($p = 0.337$). Thus, the depleted epigenomic features in the genomic regions containing elusive genes are unlikely to be false discoveries caused by a technical issue, namely, underrepresentation of the sequencing reads.

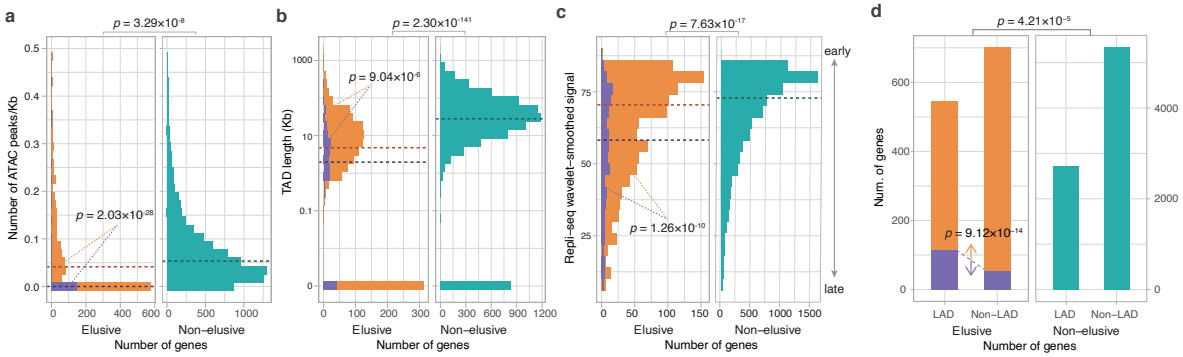


Figure 6. Epigenetic features of the elusive genes

Comparison of the distribution of ATAC-seq peak density (**a**), length of the Topologically Associating Domains (TADs) including the elusive or non-elusive genes (**b**), the replication timing indicator based on Repli-seq (**c**), and overlap with the Lamina-Associated Domains (LADs) computed from Lamin B1 ChIP-seq data. (**d**). ATAC-seq and Hi-C were performed with A549 cells, Repli-seq was performed with HepG2 cells, and Lamin B1 ChIP-seq was performed with HAP-1 cells. In the elusive gene panels, purple and orange bars indicate the elusive genes with restricted expressions ($H' < 1$; Figure 5) and those with more ubiquitous expressions ($H' \leq 1$), respectively. The results for other cells are shown in Supplementary Figures S4–S7. For each epigenetic characteristics, correction for multiple testing was performed for comparison in each cell cultures.

Discussion

Here we identified elusive genes that were lost in multiple lineages during mammalian evolution, using a comprehensive scan of gene phylogenies. To identify gene loss events, absence of evidence (i.e., missing genes caused by incomplete genome assemblies and gene annotations), should be reviewed meticulously (Deutekom et al., 2019). Additionally, gene loss might be detected erroneously because of failure in similarity searches for orthologs of rapidly evolving genes (Moyers and Zhang, 2015). In this study, we aimed to reduce these false discoveries through our multifaceted approaches (Figure 1). We selected those species with highly complete gene annotations through integration of multiple gene annotations. Using these improved gene annotations, we created orthologous groups by employing a highly sensitive homology search with MMSeq2 (Steinegger and Söding, 2017) and merged them into those identified in the Ensembl database. Furthermore, we restricted the loss events that were observed as gene absence in all species examined within all hierarchical levels of the selected taxonomic groups (Figure 1b). This absence is likely to have occurred as a gene loss in the common ancestor of the particular taxon rather than as a false discovery of gene loss in the individual species independently. Genuine continuous (e.g., telomere-to-telomere) genome assemblies are now available using modern sequencing technologies (Nurk et al., 2022). These genomic assemblies may help relieve the labor of examining for information losses, thereby facilitating the identification of genuine gene loss in any given species.

In the human genome, the elusive genes and their flanking regions harbor particular characteristics, including high GC-content and high gene density, that may have originated long before the emergence of mammals (Figure 3). Frequent

410 synonymous variations across modern humans in the elusive genes, consistent with
411 higher synonymous substitution rates between the vertebrate orthologs, suggest that the
412 genomic regions including elusive genes have been subject to rapid evolution for 500
413 million years (Figures 2 and 4). Our findings indicate that heterogeneous genomic
414 characteristics potentially affect the fate of genes at the latest period of vertebrate
415 evolution. Analyses with large numbers of germline mutations in the human genome
416 have illustrated the heterogeneity of mutation rates (Campbell and Eichler, 2013;
417 Seplyarskiy and Sunyaev, 2021; Terekhanova et al., 2017). High GC-content in the
418 elusive genes may have facilitated an elevation of the mutation rate, as observed in the
419 enrichment of rare variants in high-GC regions in the human genome (Schaibley et al.,
420 2013). In addition, some of the elusive genes appear to have retained particular
421 epigenetic marks including sparse ATAC-seq peaks, late replication timings, and
422 location within LADs (Figure 6; Supplementary Figures S4–S7); these epigenetic marks
423 are relevant to an increase in the mutation rate. Genomic regions with late replication
424 timing exhibit increased mutation rates because of their unstable structure during the S-
425 phase of the cell cycle (Koren et al., 2012; Stamatoyannopoulos et al., 2009). LADs
426 retain more G-to-A mutations because of their susceptibility to oxidative damage in the
427 nuclear periphery resulting in high levels of 8-Oxoguanine (Yoshihara et al., 2014).
428 The epigenetic marks of elusive genes are relevant to the suppression of gene
429 expression (van Steensel and Belmont, 2017), and indeed, these genes harbor weakened
430 and spatially restricted expression profiles (Figures 5–6 and S4–S7). However, the
431 genomic features associated with these epigenetic marks usually exhibit lower GC-
432 contents and reduced gene density (Gilbert et al., 2004; Rao et al., 2014; van Steensel
433 and Belmont, 2017). This discrepancy may be caused in part by a gain of local

heterochromatin accompanied with suppression of the expression of transposable elements, as observed in various eukaryotic genomes (Choi and Lee, 2020; Fiston-Lavier et al., 2007; Grewal and Jia, 2007; Rangasamy, 2013; Slotkin and Martienssen, 2007; Underwood et al., 2017). Previous analyses showed frequent heterochromatinization of the human genomic regions where KRAB zinc finger genes colocalize with L1 retrotransposons (Imbeault et al., 2017; O'Geen et al., 2007; Vogel et al., 2006). One of the genomic regions found in human chromosome region 19p12 also contains many elusive genes (Vogel et al., 2006) (Fig. 7). Closer attention to the local gene and repeat contents including repetitive elements and tandem gene clusters might facilitate our understanding of heterochromatinization in restricted genomic regions, although we excluded such gene clusters in our search for elusive genes (Figure 1).

The heterogeneous locations of elusive genes can also be interpreted from a chromosome-scale viewpoint (Figure 7a; Figure 7–figure supplement 1). Elusive genes were found in particular genomic regions including nearly all of human chromosome 19, and these regions clearly overlapped with regions of high gene density. This is consistent with our observation that the elusive genes were located in the genomic regions with higher gene density than those with the non-elusive genes. Importantly, some of these genomic regions were traced back to the microchromosomes of the ancestral gnathostomes and/or amniotes by karyotyping the ancestral genomes (Figure 7b; Figure 7–figure supplement 1). Recent studies have indicated that these microchromosomes were generated from duplicated chromosomes via allotetraploidization in early vertebrate evolution followed by rapid deletion of large parts of the chromosomal regions (Nakatani et al., 2021; Simakov et al., 2020). In

addition, vertebrate microchromosomes harbor particular genomic features including high GC-content, high gene density, and high recombination rate, some of which are concordant with genomic regions containing elusive genes (Groenen et al., 2009; International Chicken Genome Sequencing Consortium, 2004; Schield et al., 2019). This inference of ancestral karyotypes augments our observations that some elusive natures have been retained for hundreds of millions of years, and further suggests that the disparity of genomic regions has been retained for an equivalent timescale.

Finally, we note the potential evolutionary courses that facilitate the transition of gene fate from retention to loss. One possible course is a decrease in essential functions because of rapid sequence evolution in local genomic regions. The elusive genes located in those genomic regions with rapidly-evolving characteristics are likely to accumulate neutral or even moderately harmful mutations in coding regions frequently, resulting in impaired essential functions. Another factor is the spatiotemporal suppression of gene expression via epigenetic constraints. Elusive genes with restricted expressions may have limited opportunities to function, potentially leading to loss of their important roles. The extent of these evolutionary forces may have varied with time and lineages, resulting in patchy loss of elusive genes phylogenetically. Interestingly, a recent large-scale scan of *de novo* mutations in *Arabidopsis* indicates the association of mutation rates with epigenetic features and functional essentiality of genes (Monroe et al., 2022). Further investigation of the association of genes with the surrounding genomic regions in various taxa may provide a common understanding of genomic and epigenomic features that potentially alter the fate of genes. Although epigenetic features are plastic, our findings indicate that the disparities of genomic regions are reflected in the heterogeneity of evolutionary forces and have been retained for hundreds of millions of

482 years. This idea prompts us to explore evolutionary constraints on more global genomic
483 regions that are potentially associated with structural characteristics including
484 chromosomal composition and locations within the nucleus.
485

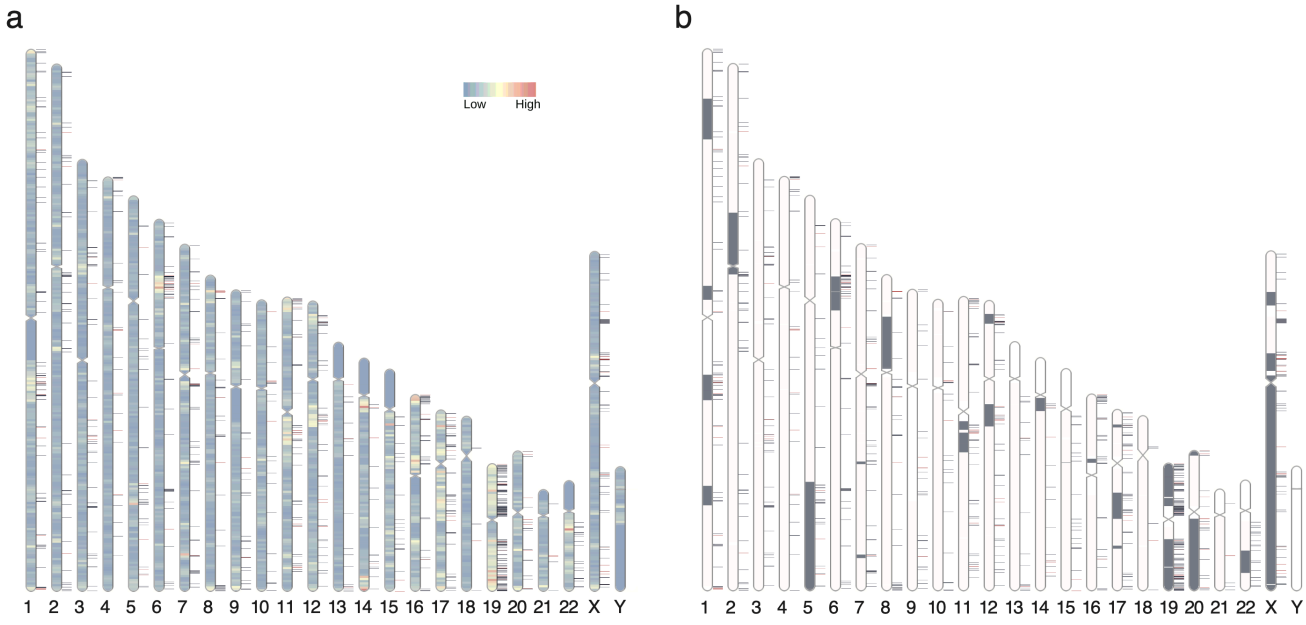


Figure 7. Chromosomal distribution of human elusive genes

Red and dark blue horizontal bars on the side of the chromosome diagram represent the location of elusive genes with restricted expression (Shannon's $H' \leq 1$) and more ubiquitous expression ($H' > 1$), respectively. (**a**) The chromosome diagrams are colored according to gene density (number of genes/Mb). (**b**) Gray regions in the diagram indicate orthologous regions of microchromosomes in the ancestors of gnathostomes (Nakatani et al., 2021). The chromosome diagrams were drawn using RIdeogram (Hao et al., 2020).

Materials and Methods

Sequence retrieval

We retrieved genome assemblies and gene annotations of 114 mammals and 132 non-mammal vertebrates from RefSeq (accessed on April 9, 2018), Ensembl release 92, and the repositories of the individual genome projects (Supplementary Table S1). Gene annotations for a single species from multiple repositories were integrated into one as follows. When gene annotations of multiple repositories were generated referring to the same version of the genome assembly, the annotation GTF files were merged with the 'cuffcompare' tool (Trapnell et al., 2012). Otherwise, translated amino acid sequences were clustered by CD-HIT v. 4.6.4 (Fu et al., 2012) with 100% sequence similarity, and the representative sequence for each cluster was retrieved by assuming that each cluster represented a single locus. Subsequently, we selected the canonical amino acid sequence for each locus: canonical peptides of the Ensembl genes were retrieved from the Ensembl database; for other resources, the longest amino acid sequence from the isoforms of a locus was chosen. The completeness of the gene annotations was performed on the gVolante web server with assessments by BUSCO v. 2 (Simão et al., 2015) by referring to the CVG (Hara et al., 2015) and BUSCO vertebrate ortholog sets. The gene annotations of mammals, birds, and ray-finned fishes that had fewer than 1% missing genes, as well as those of the other vertebrates with fewer than 3% missing genes, were selected. Exceptionally, the gene annotations of *Gavialis gangeticus* (Reptilia; CVG missing ratio 3.86%), *Paroedura picta* (Reptilia; BUSCO vertebrate ortholog missing rate 3.25%), and *Scyliorhinus torazame* (Chondrichthyes; BUSCO vertebrate ortholog missing rate 4.45%) were added. Finally, the amino acid sequence

519 set of 90 mammals and 101 non-mammalian vertebrates was subjected to t ortholog
520 clustering. We also retrieved coding nucleotide sequences of the canonical amino acid
521 sequences.
522
523 *Ortholog clustering and tree inference*
524 We retrieved gene trees of human protein-coding genes and their homologs from
525 Ensembl Gene Tree release 92. From these gene trees, we constructed an amino acid
526 sequence set of the homologs consisting of the species selected in the above section.
527 This sequence set, restricted to Ensembl sequences only, was used as the 'backbone' of
528 the ortholog set of all the selected species. In addition, we generated ortholog groups for
529 all the species used by employing OrthoFinder2 v. 2.3.3 (Emms and Kelly, 2019) based
530 on the similarity of amino acid sequences: a sequence similarity search was performed
531 using MMSeqs2 v. 2339462c06eab0bee64e4fc0ebebf7707f6e53fd (Steinegger and
532 Söding, 2017). The Ensembl and OrthoFinder ortholog sets were then merged to create
533 the united set of ortholog groups, yielding 50,768 vertebrate ortholog groups.
534 The integrated ortholog groups were then subjected to molecular phylogenetic
535 analysis. Amino acid sequences of the individual groups were aligned with MAFFT v.
536 7.402 (Katoh and Standley, 2013), and ambiguous alignment sites were removed with
537 trimAl v1.4 (Capella-Gutiérrez et al., 2009). Phylogenetic trees were inferred with IQ-
538 Tree v. 1.6.6 (Nguyen et al., 2015) by selecting the optimal amino acid substitution
539 model with ModelFinder (Kalyaanamoorthy et al., 2017) implemented in the IQ-Tree
540 tool for each sequence alignment. In the inferred phylogenetic trees, ambiguously
541 bifurcated nodes—those with branch lengths less than 0.0025—were collapsed into a
542 multifurcational node by the 'di2multi' function implemented in ape v. 5.5 (Paradis and

543 Schliep, 2019). The trees were then rooted with the automatic rooting function
544 'get_age_balanced_outgroup' implemented in ete3 v. 3.1.1 (Huerta-Cepas et al., 2016)
545 to minimize any discrepancy of tree topologies with the taxonomic hierarchy of the
546 species included.
547
548 *Identification of elusive genes in the human genome*
549 For the individual trees, orthologs of the human genes were detected by the
550 'get_my_evol_events' function in ete3 (Huerta-Cepas et al., 2007). This function
551 inferred gene duplication nodes in the rooted trees, resulting in separation of the trees
552 into 17,495 subtrees of mammalian ortholog groups containing human genes. The
553 ortholog information was referenced to extract the species with no orthologs to
554 humans . This absence was further assessed by the ortholog annotation of human genes
555 in the Ensembl Gene Tree database.
556 We selected taxonomic groups for the individual mammalian ortholog groups in
557 which the orthologs were missing in all the species examined (Supplementary Table
558 S1). We restricted our study to gene losses that were likely to have occurred in the
559 common ancestor of particular taxonomic groups, rather than those arising from the
560 incompleteness of gene annotations. When a gene was missing in all the taxonomic
561 groups in the same hierarchy, we considered that the gene was lost in the common
562 ancestor of these groups. Finally, we found 1,233 human genes belonging to the
563 ortholog groups that were absent in two or more taxonomic groups and defined them as
564 elusive genes. We further selected 1,081 elusive genes that harbored three or fewer
565 mammalian paralogs for the following analyses. Similarly, we extracted 8,050 human
566 genes whose orthologs were found in all the mammalian species examined and defined

567 them as non-elusive genes. Because these elusive and non-elusive genes were identified
568 in the GRCh38 human genome, we performed the following analyses referring to that
569 assembly.
570

571 *Extraction of genomic and molecular evolutionary characteristics*

572 We calculated the GC content of a gene by using its genomic region including introns
573 and untranslated regions (UTRs). To calculate individual gene densities, we extracted
574 genomic regions containing the genes and their flanking three genes at both ends and
575 divided them by seven. The orthologs of the elusive and non-elusive genes were
576 retrieved from the aforementioned gene trees. Amino acid sequence alignment of the
577 human and the ortholog genes was performed using MAFFT. Nucleotide sequence
578 alignments of the coding regions were generated by 'back-translation' of the amino acid
579 sequence alignments by trimAl, simultaneously removing ambiguous alignment sites.
580 By employing coding nucleotide sequence alignments, numbers of synonymous and
581 non-synonymous substitutions per site were computed using PAML v. 4.9a (Yang,
582 2007).
583

584 *Multiomics analysis*

585 Common and rare SNVs of the human populations were retrieved from dbSNP release
586 147 (Sherry et al., 2001), and human CNVs were obtained from the Database of
587 Genomic Variants (DGV) release 2016-08-31 (MacDonald et al., 2014). The CNVs
588 were classified into duplication and deletion variants, according to the annotation in
589 DGV. The density of these variants in a gene was computed by dividing the number of

590 variants identified in a gene region by its sequence length. Z-scores, indices of the
591 tolerance against mutations, of synonymous, missense, and loss-of-function mutations
592 of the individual genes were retrieved from gnomAD v. 2.1.1 (Karczewski et al., 2021).
593 Gene expression quantifications of adult and fetal tissues were retrieved from
594 public databases. Expression profiles of adult tissues were obtained from the GTEx
595 database v. 8 (GTEx Consortium, 2020), computed by averaging TPM values across
596 individuals. Expression profiles of fetal tissues were obtained from the Developmental
597 Single Cell Atlas of gene Regulation and Expression (Descartes) portal (Cao et al.,
598 2020), by calculating averaged TPM values of single cells. The maximum TPM values
599 of the individual genes among the tissues were taken as the representative gene
600 expression levels. As a proxy of the spatial diversity of gene expression, Shannon's
601 species diversity index (H' values) were computed for the individual genes using the
602 following equation:

603 $$H_i' = -\sum_{k=1}^{R} p_{i,k} \ln p_{i,k}$$

604 where H_i' represents the Shannon's index of ith gene in the list of the human genes, $p_{i,k}$
605 represents the proportion of the TPM values of the ith gene in the kth tissues/cell types,
606 and R denotes the total number of tissues/cell types examined.
607 The ATAC-seq peaks and TAD boundaries of the human primary cells and
608 culture strains were retrieved from the ENCODE 3 repository (Accession ID listed in
609 Table S3) (ENCODE Project Consortium, 2012). Wavelet-smoothed signals of the
610 ENCODE Repli-seq data were obtained from the UCSC genome browser (Hansen et al.,
611 2010). The 20 kb bin-associated domains of LAD-seq that employed Lamin B1

antibodies (van Schaik et al., 2020) were retrieved from the 4D Nucleome Data Portal (https://data.4dnucleome.org/publications/f1218a92-1f37-4519-85d6-ccedd5f7ad39).

Code availability

The scripts for inferring gene presence and absence from gene tree was deposited in GitHub (https://github.com/yuichiroharajpn/ElusiveGenes).

Statistical tests

Comparisons of the genomic characteristics between the elusive and non-elusive genes were tested statistically with the nonparametric Mann–Whitney *U* test and Fisher's exact test implemented in R. Correction of multiple testing was performed using the Benjamini–Hochberg false discovery rate (FDR) approach.

Supplementary Figures

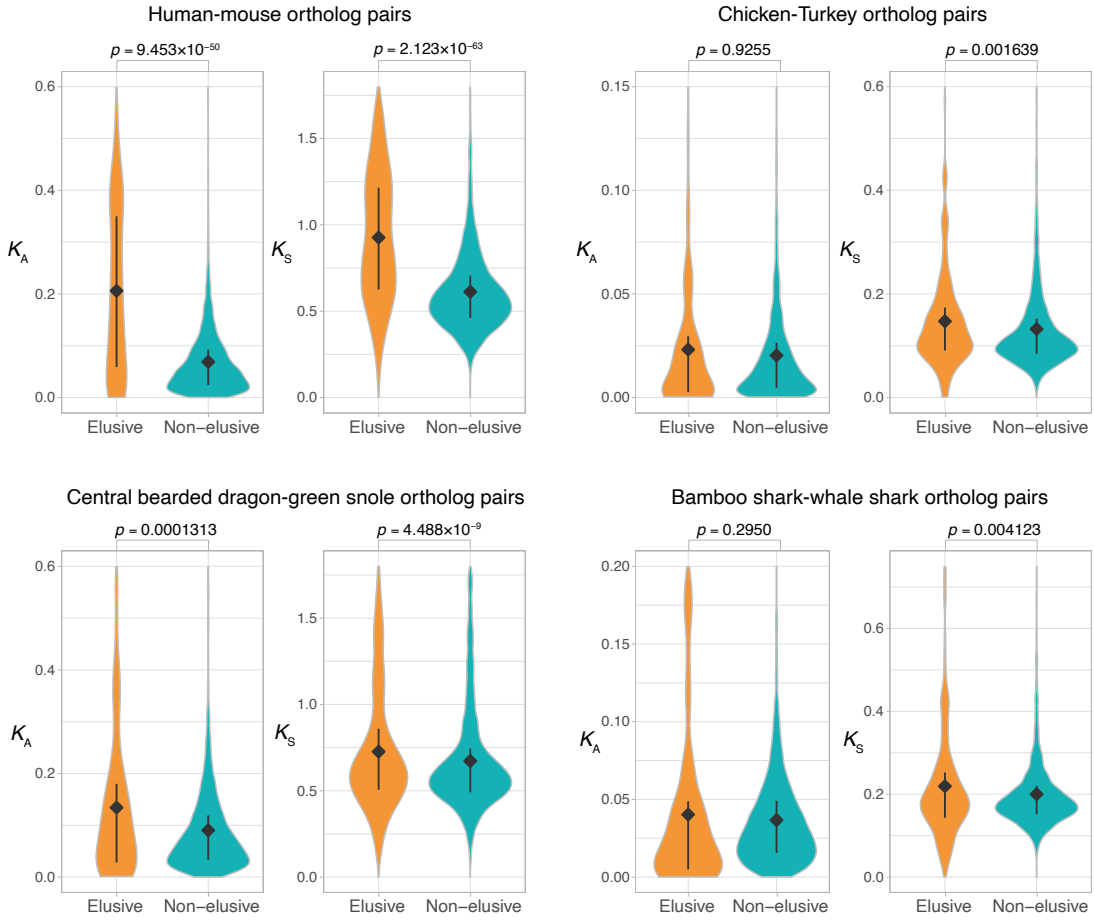
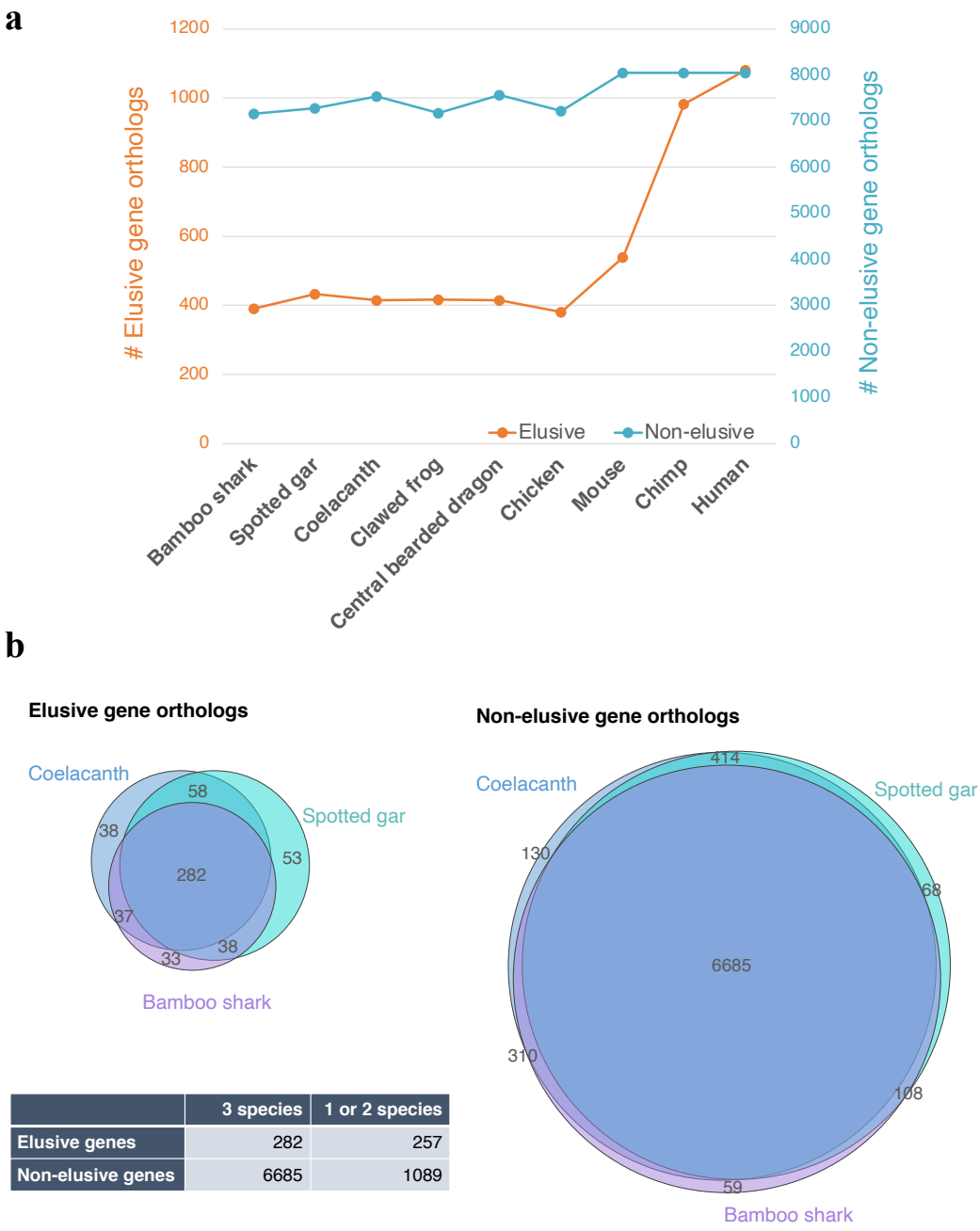


Figure 2–figure supplement 1. Comparison of K_A and K_S values between orthologs of the elusive and non-elusive genes

Distributions of K_A and K_S values between the orthologs of human elusive and non-elusive genes of closely related vertebrates. Correction for multiple testing was performed for comparison in each species pair.



	3 species	1 or 2 species
Elusive genes	282	257
Non-elusive genes	6685	1089

634

635 **Figure 3–figure supplement 1. Asymmetric ortholog retention across the**

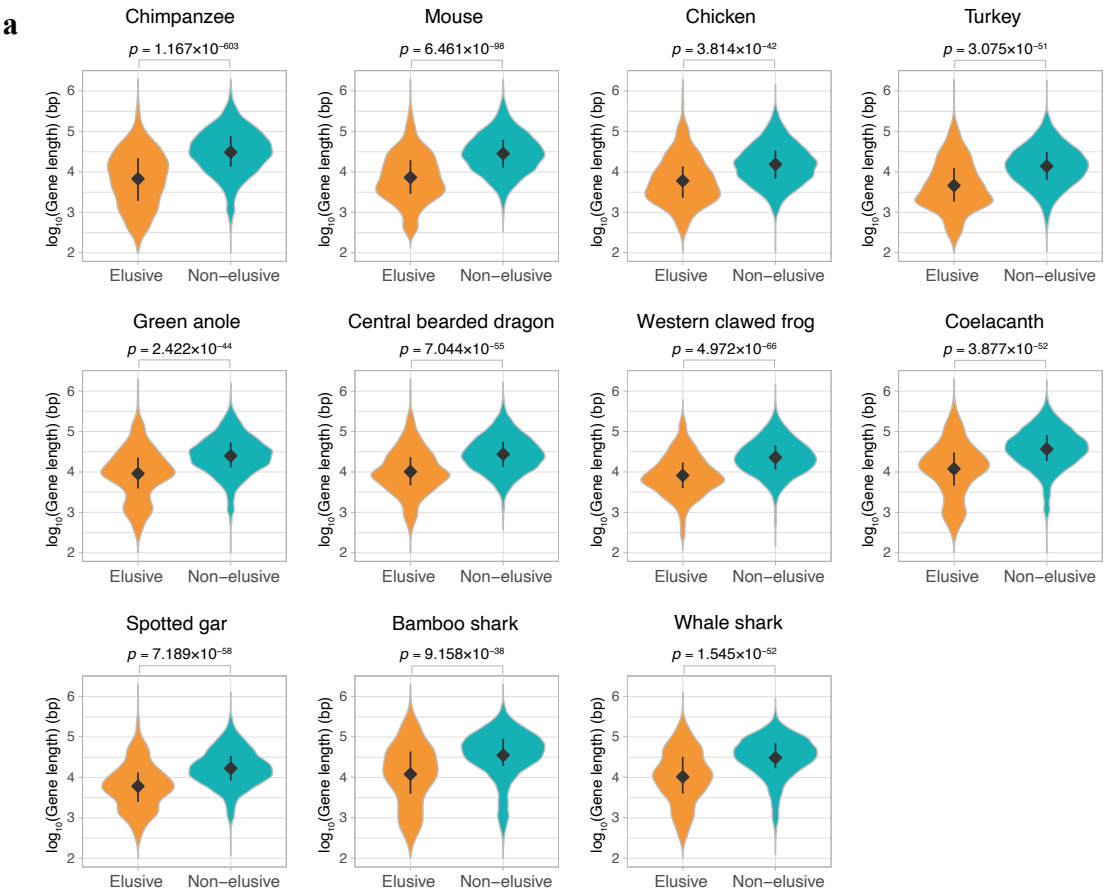
636 **vertebrates**

637 Number of retained orthologs of the human elusive and non-elusive genes (**a**) and the

638 overlaps of the retained orthologs across three vertebrates distantly related to modern

639 humans (**b**). The *p*-value of the 2 × 2 contingency table given by Fisher's exact test is

640 4.5×10^{-71}.



641

642 **Figure 3–figure supplement 2. Genomic characteristics of the orthologs of elusive**

643 **and non-elusive genes**

644 Distribution of (**a**) gene length and (**b**) GC-content of the orthologs of the human

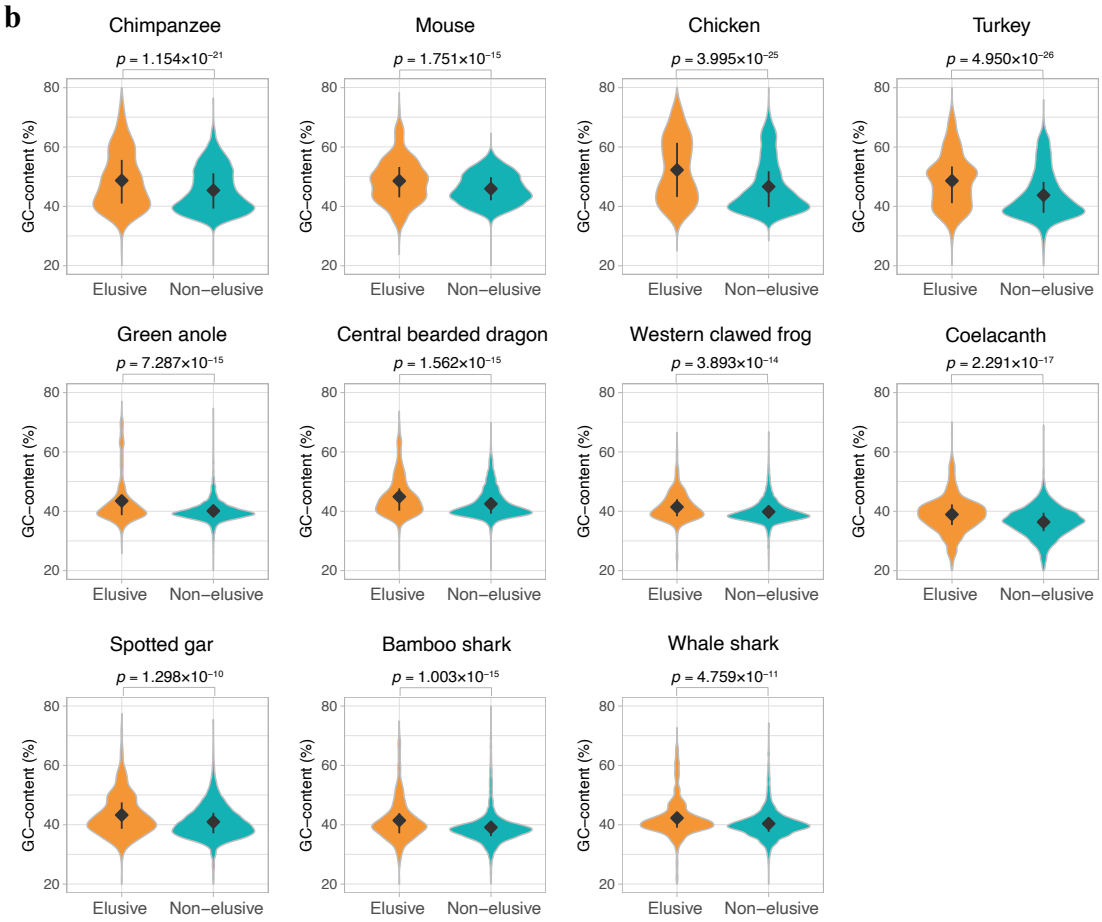
645 elusive and non-elusive genes and (**c**) distribution of the gene density of the genomic

646 regions where the orthologs of the human elusive and non-elusive genes are located. For

647 each genomic characteristics, correction for multiple testing was performed for

648 comparison in each species.

649

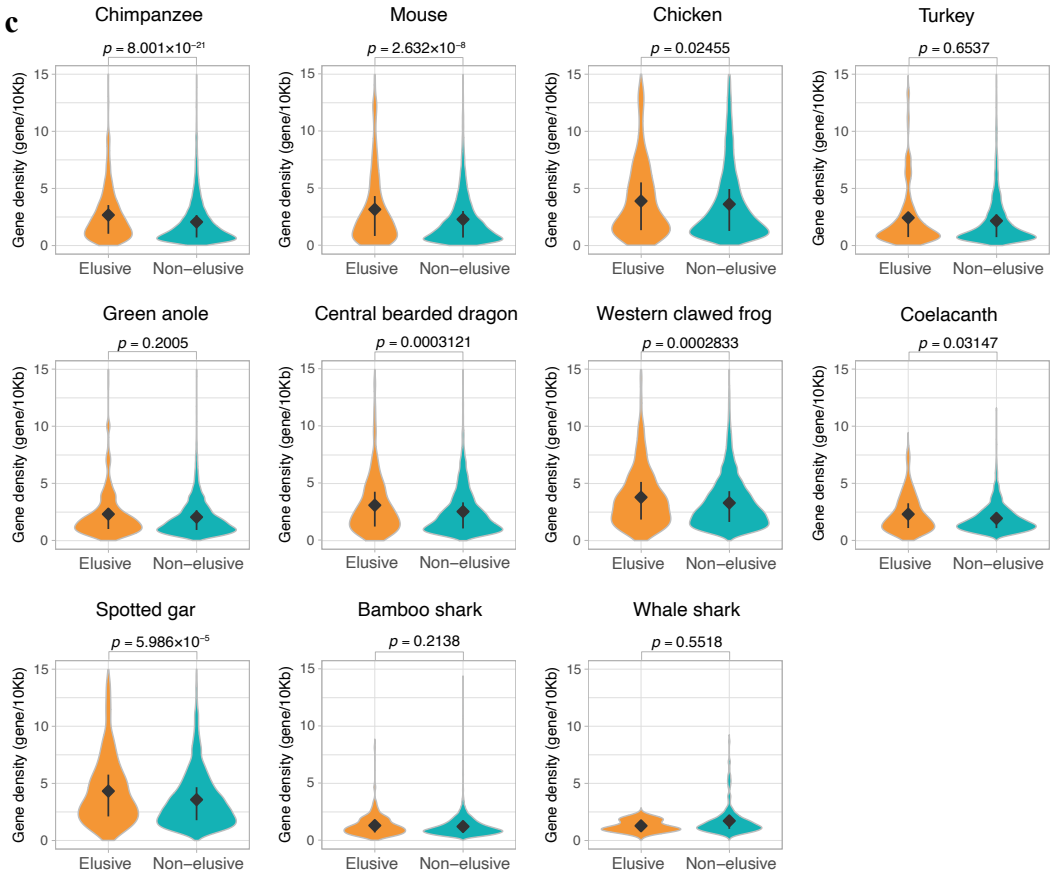


650

651

652 **Figure 3–figure supplement 2 (continued). Genomic characteristics of the**

653 **orthologs of elusive and non-elusive genes**

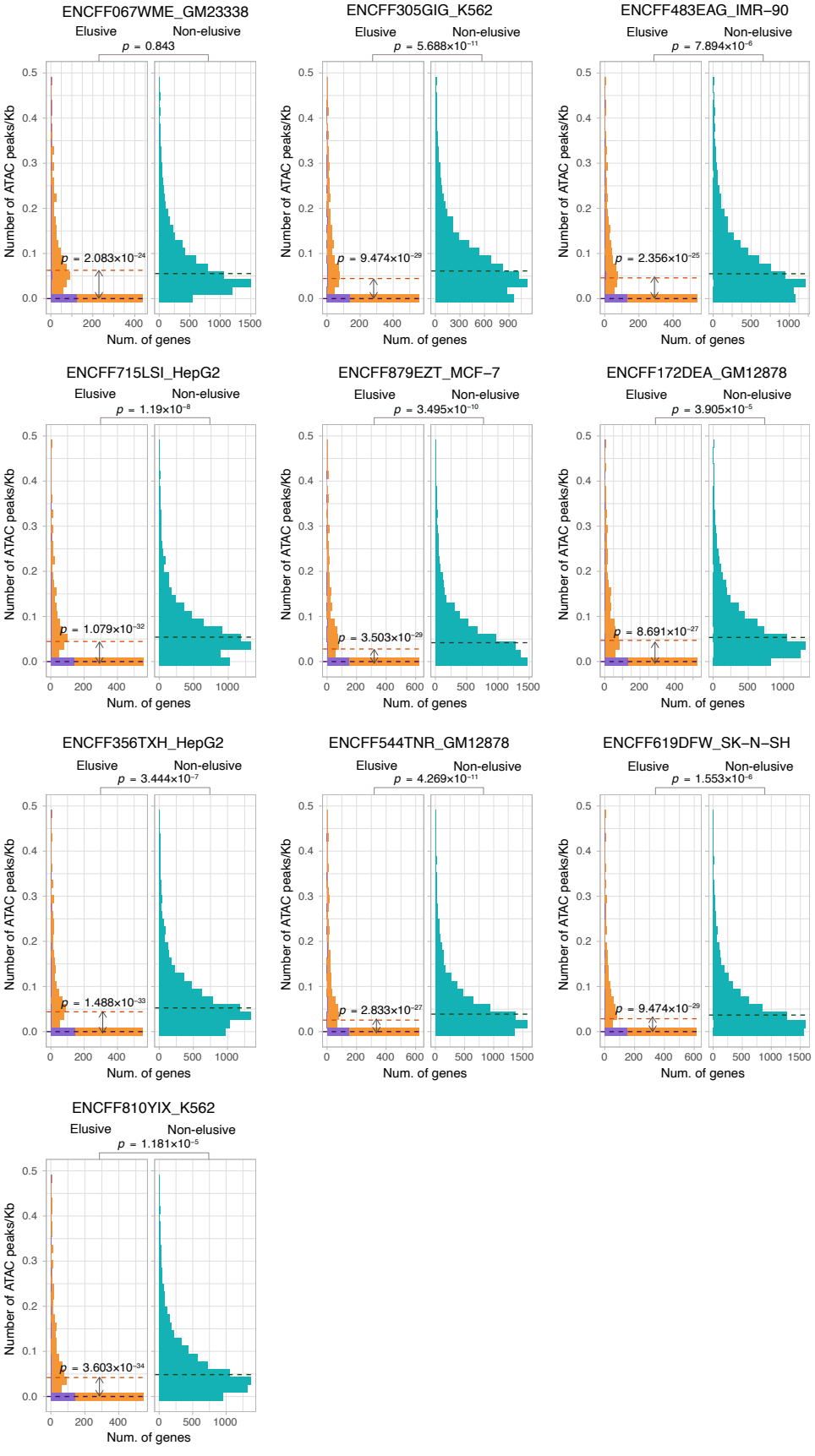


654

655 **Figure 3–figure supplement 2 (continued). Genomic characteristics of the**

656 **orthologs of elusive and non-elusive genes**

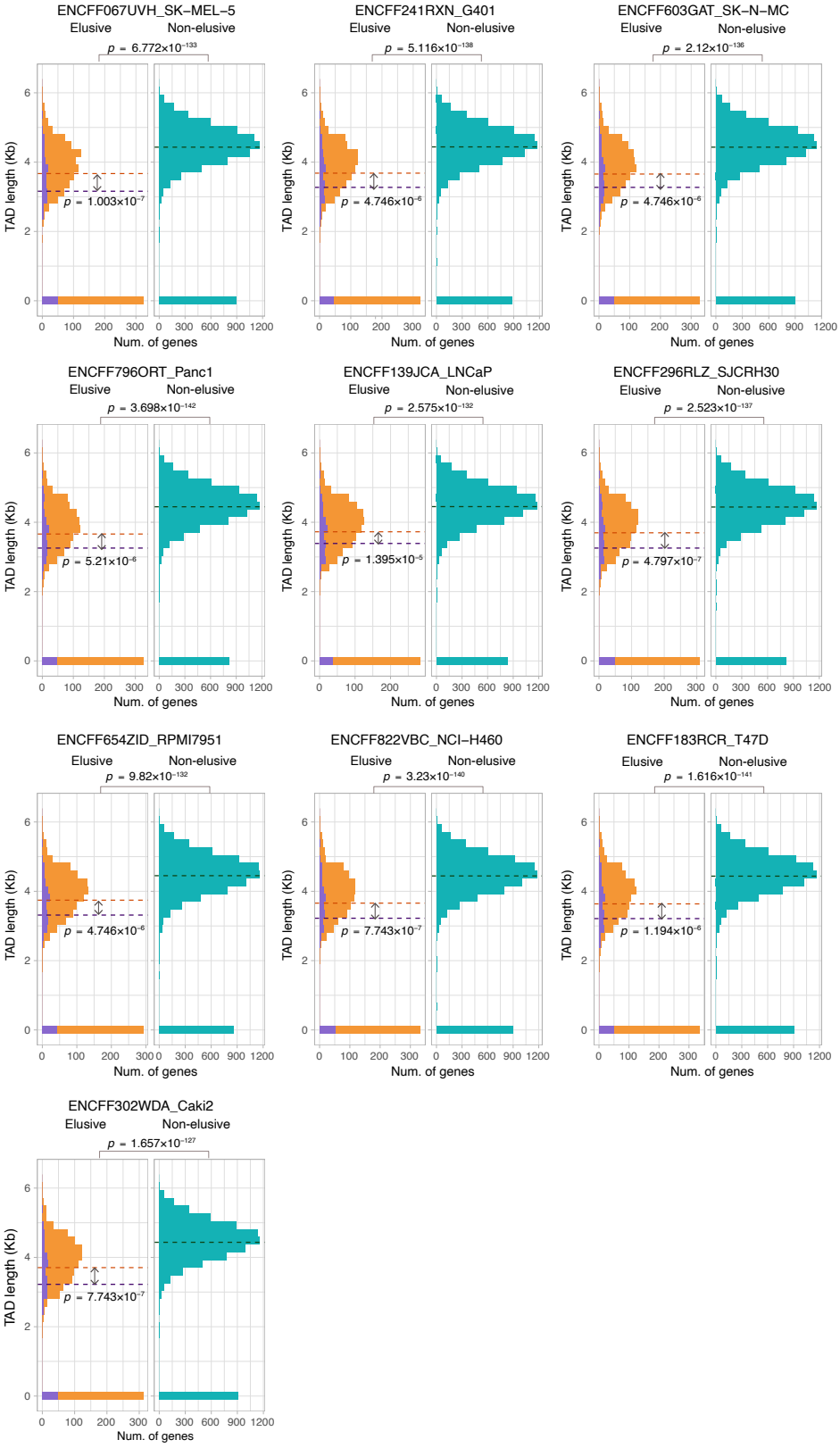
657



658

659 **Figure 6–figure supplement 1. ATAC-seq peak density of the elusive and non-**
660 **elusive gene regions**

661 Comparison of the distribution of ATAC-seq peak density between the elusive and non-
662 elusive genes across multiple cell types. In the elusive gene panels, purple and orange
663 bars indicate the elusive genes with restricted expressions ($H' < 1$; Figure 5) and those
664 with more ubiquitous expressions ($H' \leq 1$), respectively. Correction for multiple testing
665 was performed for comparison in each cell cultures.



666

667 **Figure 6–figure supplement 2. Sequence lengths of the topologically associating**

668 **domains (TADs) containing elusive or non-elusive genes**

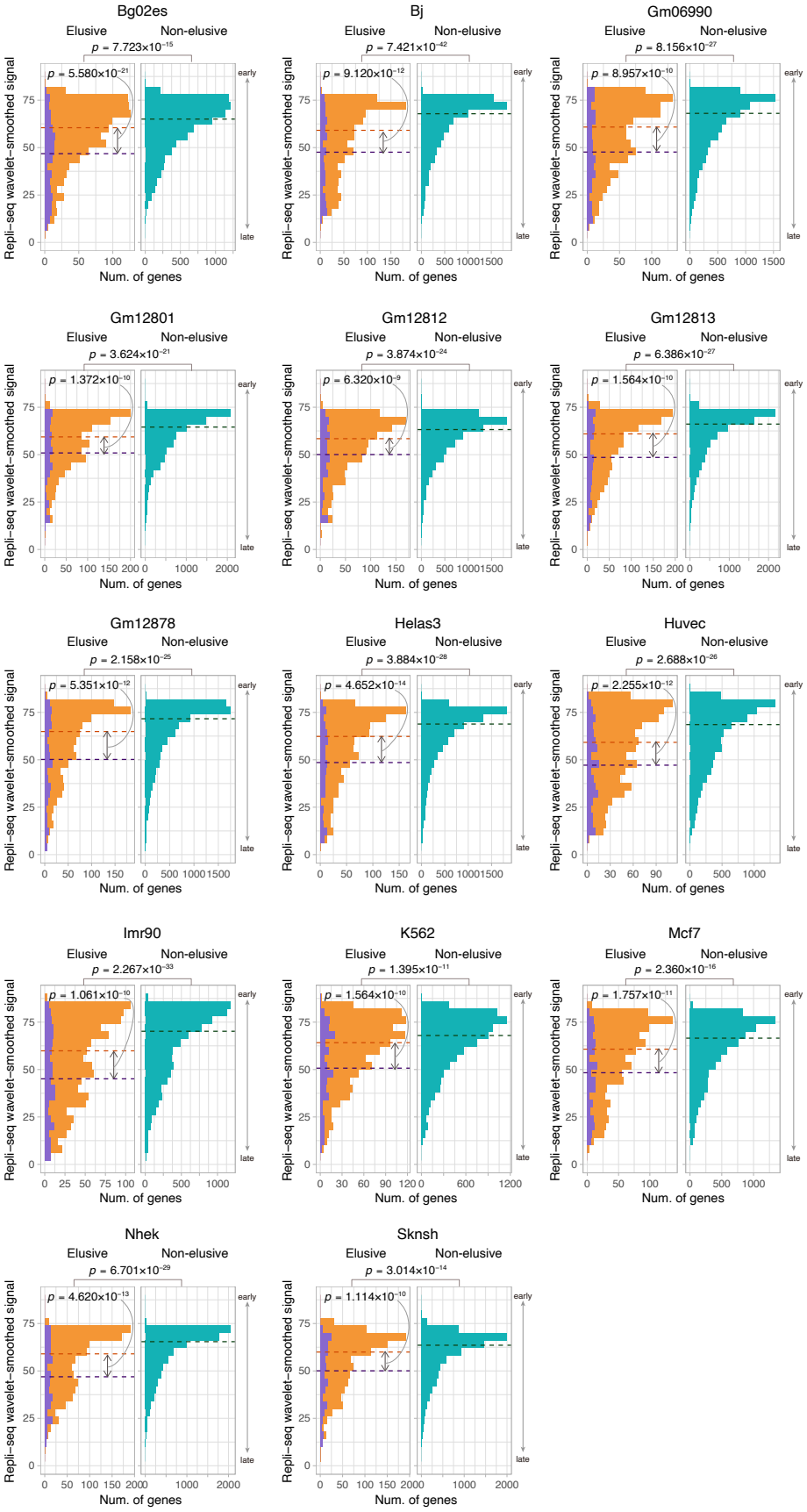
669 Comparison of the distribution of length of TADs including the elusive or non-elusive

670 genes across multiple cell types. In the elusive gene panels, purple and orange bars

671 indicate the elusive genes with restricted expressions ($H' < 1$; Figure 5) and those with

672 more ubiquitous expressions ($H' \leq 1$), respectively. Correction for multiple testing was

673 performed for comparison in each cell cultures.



674

675 **Figure 6–figure supplement 3. Comparison of the replication timing indicator**
676 **based on Repli-seq between the elusive and non-elusive genes.**
677 Comparison of the distribution of replication timing indicator based on Repli-seq
678 between the elusive and non-elusive genes across multiple cell types. In the elusive gene
679 panels, purple and orange bars indicate elusive genes with restricted expressions ($H' <$
680 1; Figure 5) and those with more ubiquitous expressions ($H' \leq 1$), respectively.
681 Correction for multiple testing was performed for comparison in each cell cultures.

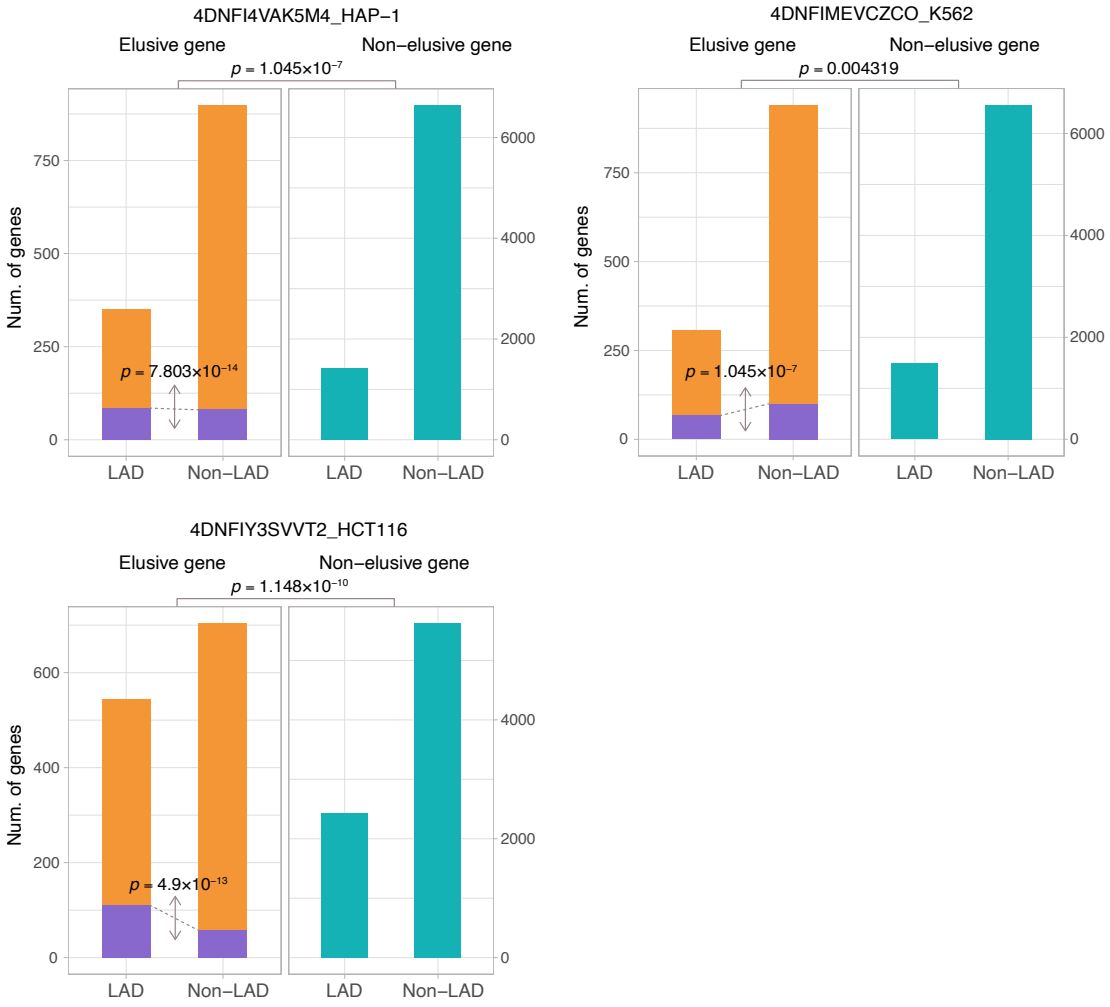


Figure 6–figure supplement 4. The fraction of elusive and non-elusive genes that overlap with Lamina-Associated Domains (LADs)

Comparison of frequency of overlap with LADs computed from Lamin B1 ChIP-seq data between the elusive and non-elusive genes across multiple data. In the elusive gene panels, purple and orange bars indicate elusive genes with restricted expressions ($H' < 1$; Figure 5) and those with more ubiquitous expressions ($H' \leq 1$), respectively. The results for other cells are shown in Figures S4–S7.

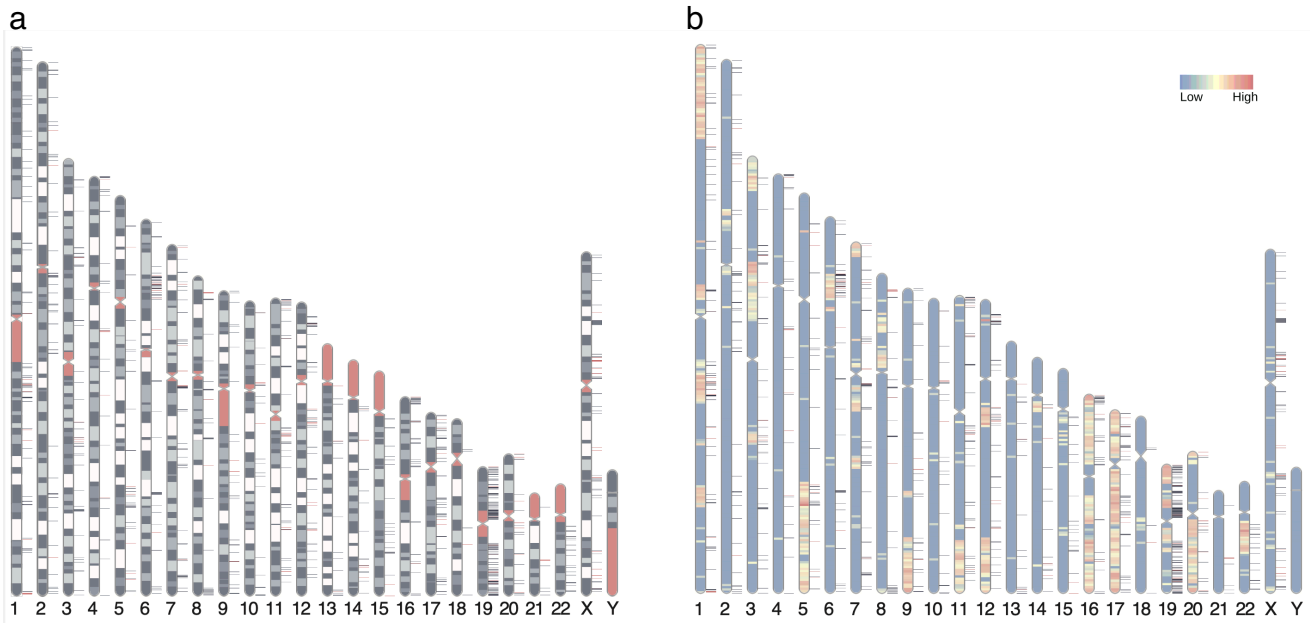


Figure 7–figure supplement 1. Distribution of elusive genes across human chromosomes

Red and dark blue horizontal bars on the side of the chromosome diagram represent the location of elusive genes with restricted expression (Shannon's $H' \leq 1$) and more ubiquitous expression ($H' > 1$), respectively. (**a**) Karyotypes are shown by G-banding. Red regions indicate centromeres, acrocentric regions, and variable-length regions. (**b**) The chromosome diagrams are colored according to the density of the genes that harbor chicken orthologs in microchromosomes (number of genes/Mb). The chromosome diagrams were drawn using RIdeogram (GTEx Consortium, 2020).

Acknowledgements

We thank Dr. Yoichiro Nakatani for providing the information of orthologous regions of ancestral chromosomes in the human genome. This work was supported by RIKEN to S.K., JSPS KAKENHI Grant Number 20H03269 to S.K. and 21K06132 to Y.H., and Mochida Memorial Foundation for Medical and Pharmaceutical Research to Y.H. Computations were partially performed on the NIG supercomputer at ROIS National Institute of Genetics.

Competing interests

The authors declare that they have no competing interests.

References

Albalat R, Cañestro C. 2016. Evolution by gene loss. *Nat Rev Genet* **17**:379–391.

Bartha I, di Iulio J, Venter JC, Telenti A. 2018. Human gene essentiality. *Nat Rev Genet* **19**:51–62.

Bernardi G, Bernardi G. 1986. Compositional constraints and genome evolution. *J Mol Evol* **24**:1–11.

Blanc G, Agarkova I, Grimwood J, Kuo A, Brueggeman A, Dunigan DD, Gurnon J, Ladunga I, Lindquist E, Lucas S, Pangilinan J, Pröschold T, Salamov A, Schmutz J, Weeks D, Yamada T, Lomsadze A, Borodovsky M, Claverie J-M, Grigoriev IV, Van Etten JL. 2012. The genome of the polar eukaryotic microalga Coccomyxa subellipsoidea reveals traits of cold adaptation. *Genome Biol* **13**:R39.

Botero-Castro F, Figuet E, Tilak M-K, Nabholz B, Galtier N. 2017. Avian Genomes Revisited: Hidden Genes Uncovered and the Rates versus Traits Paradox in Birds. *Mol Biol Evol* **34**:3123–3131.

Byrne KP, Wolfe KH. 2007. Consistent patterns of rate asymmetry and gene loss indicate widespread neofunctionalization of yeast genes after whole-genome duplication. *Genetics* **175**:1341–1350.

Campbell CD, Eichler EE. 2013. Properties and rates of germline mutations in humans. *Trends Genet* **29**:575–584.

Cao J, O'Day DR, Pliner HA, Kingsley PD, Deng M, Daza RM, Zager MA, Aldinger KA, Blecher-Gonen R, Zhang F, Spielmann M, Palis J, Doherty D, Steemers FJ, Glass IA, Trapnell C, Shendure J. 2020. A human cell atlas of fetal gene expression. *Science* **370**:eaba7721.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**:1972–1973.

Choi JY, Lee YCG. 2020. Double-edged sword: The evolutionary consequences of the epigenetic silencing of transposable elements. *PLoS Genet* **16**:e1008872.

Cohen N, Dagan T, Stone L, Graur D. 2005. GC composition of the human genome: in search of isochores. *Mol Biol Evol* **22**:1260–1272.

Cortez D, Marin R, Toledo-Flores D, Froidevaux L, Liechti A, Waters PD, Grützner F, Kaessmann H. 2014. Origins and functional evolution of Y chromosomes across mammals. *Nature* **508**:488–493.

745 Debatisse M, Le Tallec B, Letessier A, Dutrillaux B, Brison O. 2012. Common fragile
746 sites: mechanisms of instability revisited. *Trends Genet* **28**:22–32.
747 Deutekom ES, Vosseberg J, van Dam TJP, Snel B. 2019. Measuring the impact of gene
748 prediction on gene loss estimates in Eukaryotes by quantifying falsely inferred
749 absences. *PLoS Comput Biol* **15**:e1007301.
750 Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for
751 comparative genomics. *Genome Biol* **20**:238.
752 ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in
753 the human genome. *Nature* **489**:57–74.
754 Fernández R, Gabaldón T. 2020. Gene gain and loss across the metazoan tree of life.
755 *Nat Ecol Evol* **4**:524–533.
756 Fiston-Lavier A-S, Anxolabehere D, Quesneville H. 2007. A model of segmental
757 duplication formation in Drosophila melanogaster. *Genome Res* **17**:1458–1470.
758 Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-
759 generation sequencing data. *Bioinformatics* **28**:3150–3152.
760 Giaever G, Nislow C. 2014. The yeast deletion collection: a decade of functional
761 genomics. *Genetics* **197**:451–465.
762 Gilbert N, Boyle S, Fiegler H, Woodfine K, Carter NP, Bickmore WA. 2004.
763 Chromatin architecture of the human genome: gene-rich domains are enriched in
764 open chromatin fibers. *Cell* **118**:555–566.
765 Grewal SIS, Jia S. 2007. Heterochromatin revisited. *Nat Rev Genet* **8**:35–46.
766 Groenen MAM, Wahlberg P, Foglio M, Cheng HH, Megens H-J, Crooijmans RPMA,
767 Besnier F, Lathrop M, Muir WM, Wong GK-S, Gut I, Andersson L. 2009. A
768 high-density SNP-based linkage map of the chicken genome reveals sequence
769 features correlated with recombination rate. *Genome Res* **19**:510–519.
770 GTEx Consortium. 2020. The GTEx Consortium atlas of genetic regulatory effects
771 across human tissues. *Science* **369**:1318–1330.
772 Guijarro-Clarke C, Holland PWH, Paps J. 2020. Widespread patterns of gene loss in the
773 evolution of the animal kingdom. *Nat Ecol Evol* **4**:519–523.
774 Hansen RS, Thomas S, Sandstrom R, Canfield TK, Thurman RE, Weaver M, Dorschner
775 MO, Gartler SM, Stamatoyannopoulos JA. 2010. Sequencing newly replicated
776 DNA reveals widespread plasticity in human replication timing. *Proc Natl Acad*
777 *Sci U S A* **107**:139–144.
778 Hao Z, Lv D, Ge Y, Shi J, Weijers D, Yu G, Chen J. 2020. RIdeogram: drawing SVG
779 graphics to visualize and map genome-wide data on the idiograms. *PeerJ*
780 *Comput Sci* **6**:e251.

781 Hara Y, Takeuchi M, Kageyama Y, Tatsumi K, Hibi M, Kiyonari H, Kuraku S. 2018a.
782 Madagascar ground gecko genome analysis characterizes asymmetric fates of
783 duplicated genes. *BMC Biol* **16**:40.
784 Hara Y, Tatsumi K, Yoshida M, Kajikawa E, Kiyonari H, Kuraku S. 2015. Optimizing
785 and benchmarking de novo transcriptome sequencing: from library preparation
786 to assembly evaluation. *BMC Genomics* **16**:977.
787 Hara Y, Yamaguchi K, Onimaru K, Kadota M, Koyanagi M, Keeley SD, Tatsumi K,
788 Tanaka K, Motone F, Kageyama Y, Nozu R, Adachi N, Nishimura O,
789 Nakagawa R, Tanegashima C, Kiyatake I, Matsumoto R, Murakumo K, Nishida
790 K, Terakita A, Kuratani S, Sato K, Hyodo S, Kuraku S. 2018b. Shark genomes
791 provide insights into elasmobranch evolution and the origin of vertebrates. *Nat*
792 *Ecol Evol* **2**:1761–1771.
793 Helmrich A, Stout-Weider K, Hermann K, Schrock E, Heiden T. 2006. Common fragile
794 sites are conserved features of human and mouse chromosomes and relate to
795 large active genes. *Genome Res* **16**:1222–1230.
796 Hirsh AE, Fraser HB. 2001. Protein dispensability and rate of evolution. *Nature*
797 **411**:1046–1049.
798 Huerta-Cepas J, Dopazo H, Dopazo J, Gabaldón T. 2007. The human phylome. *Genome*
799 *Biol* **8**:R109.
800 Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: Reconstruction, Analysis, and
801 Visualization of Phylogenomic Data. *Mol Biol Evol* **33**:1635–1638.
802 Hughes JF, Skaletsky H, Brown LG, Pyntikova T, Graves T, Fulton RS, Dugan S, Ding
803 Y, Buhay CJ, Kremitzki C, Wang Q, Shen H, Holder M, Villasana D, Nazareth
804 LV, Cree A, Courtney L, Veizer J, Kotkiewicz H, Cho T-J, Koutseva N, Rozen
805 S, Muzny DM, Warren WC, Gibbs RA, Wilson RK, Page DC. 2012. Strict
806 evolutionary conservation followed rapid gene loss on human and rhesus Y
807 chromosomes. *Nature* **483**:82–86.
808 Imbeault M, Helleboid P-Y, Trono D. 2017. KRAB zinc-finger proteins contribute to
809 the evolution of gene regulatory networks. *Nature* **543**:550–554.
810 International Chicken Genome Sequencing Consortium. 2004. Sequence and
811 comparative analysis of the chicken genome provide unique perspectives on
812 vertebrate evolution. *Nature* **432**:695–716.
813 Jordan IK, Rogozin IB, Wolf YI, Koonin EV. 2002. Essential genes are more
814 evolutionarily conserved than are nonessential genes in bacteria. *Genome Res*
815 **12**:962–968.

816 Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017.
817 ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat*
818 *Methods* **14**:587–589.
819 Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL,
820 Laricchia KM, Ganna A, Birnbaum DP, Gauthier LD, Brand H, Solomonson M,
821 Watts NA, Rhodes D, Singer-Berk M, England EM, Seaby EG, Kosmicki JA,
822 Walters RK, Tashman K, Farjoun Y, Banks E, Poterba T, Wang A, Seed C,
823 Whiffin N, Chong JX, Samocha KE, Pierce-Hoffman E, Zappala Z, O'Donnell-
824 Luria AH, Minikel EV, Weisburd B, Lek M, Ware JS, Vittal C, Armean IM,
825 Bergelson L, Cibulskis K, Connolly KM, Covarrubias M, Donnelly S, Ferriera
826 S, Gabriel S, Gentry J, Gupta N, Jeandet T, Kaplan D, Llanwarne C, Munshi R,
827 Novod S, Petrillo N, Roazen D, Ruano-Rubio V, Saltzman A, Schleicher M,
828 Soto J, Tibbetts K, Tolonen C, Wade G, Talkowski ME, Genome Aggregation
829 Database Consortium, Neale BM, Daly MJ, MacArthur DG. 2021. Author
830 Correction: The mutational constraint spectrum quantified from variation in
831 141,456 humans. *Nature* **590**:E53.
832 Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7:
833 improvements in performance and usability. *Mol Biol Evol* **30**:772–780.
834 Katzman S, Capra JA, Haussler D, Pollard KS. 2011. Ongoing GC-biased evolution is
835 widespread in the human genome and enriched near recombination hot spots.
836 *Genome Biol Evol* **3**:614–626.
837 Koren A, Polak P, Nemesh J, Michaelson JJ, Sebat J, Sunyaev SR, McCarroll SA. 2012.
838 Differential relationship of DNA replication timing to different forms of human
839 mutation and variation. *Am J Hum Genet* **91**:1033–1040.
840 Korenberg JR, Rykowski MC. 1988. Human genome organization: Alu, lines, and the
841 molecular structure of metaphase chromosome bands. *Cell* **53**:391–400.
842 Krylov DM, Wolf YI, Rogozin IB, Koonin EV. 2003. Gene loss, protein sequence
843 divergence, gene dispensability, expression level, and interactivity are correlated
844 in eukaryotic evolution. *Genome Res* **13**:2229–2235.
845 Lewin TD, Royall AH, Holland PWH. 2021. Dynamic Molecular Evolution of
846 Mammalian Homeobox Genes: Duplication, Loss, Divergence and Gene
847 Conversion Sculpt PRD Class Repertoires. *J Mol Evol* **89**:396–414.
848 Liu G, Yong MYJ, Yurieva M, Srinivasan KG, Liu J, Lim JSY, Poidinger M, Wright
849 GD, Zolezzi F, Choi H, Pavelka N, Rancati G. 2015. Gene Essentiality Is a
850 Quantitative Property Linked to Cellular Evolvability. *Cell* **163**:1388–1399.

MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW. 2014. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res* **42**:D986-92.

Maclean CJ, Metzger BPH, Yang J-R, Ho W-C, Moyers B, Zhang J. 2017. Deciphering the Genic Basis of Yeast Fitness Variation by Simultaneous Forward and Reverse Genetics. *Mol Biol Evol* **34**:2486–2502.

Maeso I, Dunwell TL, Wyatt CDR, Marlétaz F, Vető B, Bernal JA, Quah S, Irimia M, Holland PWH. 2016. Evolutionary origin and functional divergence of totipotent cell homeobox genes in eutherian mammals. *BMC Biol* **14**:45.

Medstrand P, van de Lagemaat LN, Mager DL. 2002. Retroelement distributions in the human genome: variations associated with age and proximity to genes. *Genome Res* **12**:1483–1495.

Miyata T, Yasunaga T, Nishida T. 1980. Nucleotide sequence divergence and functional constraint in mRNA evolution. *Proc Natl Acad Sci U S A* **77**:7328–7332.

Monroe JG, Srikant T, Carbonell-Bejerano P, Becker C, Lensink M, Exposito-Alonso M, Klein M, Hildebrandt J, Neumann M, Kliebenstein D, Weng M-L, Imbert E, Ågren J, Rutter MT, Fenster CB, Weigel D. 2022. Mutation bias reflects natural selection in Arabidopsis thaliana. *Nature* **602**:101–105.

Moyers BA, Zhang J. 2015. Phylostratigraphic bias creates spurious patterns of genome evolution. *Mol Biol Evol* **32**:258–267.

Nakatani Y, Shingate P, Ravi V, Pillai NE, Prasad A, McLysaght A, Venkatesh B. 2021. Reconstruction of proto-vertebrate, proto-cyclostome and proto-gnathostome genomes provides new insights into early vertebrate evolution. *Nat Commun* **12**:4489.

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

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, Aganezov S, Hoyt SJ, Diekhans M, Logsdon GA, Alonge M, Antonarakis SE, Borchers M, Bouffard GG, Brooks SY, Caldas GV, Chen N-C, Cheng H, Chin C-S, Chow W, de Lima LG, Dishuck PC, Durbin R, Dvorkina T, Fiddes IT, Formenti G, Fulton RS, Fungtammasan A, Garrison E, Grady PGS, Graves-Lindsay TA, Hall IM, Hansen NF, Hartley GA, Haukness M, Howe K, Hunkapiller MW, Jain C, Jain M, Jarvis ED, Kerpedjiev P, Kirsche M, Kolmogorov M, Korlach J, Kremitzki M, Li H, Maduro VV, Marschall T, McCartney AM, McDaniel J, Miller DE, Mullikin JC,

887 Myers EW, Olson ND, Paten B, Peluso P, Pevzner PA, Porubsky D, Potapova T,
888 Rogaev EI, Rosenfeld JA, Salzberg SL, Schneider VA, Sedlazeck FJ, Shafin K,
889 Shew CJ, Shumate A, Sims Y, Smit AFA, Soto DC, Sović I, Storer JM, Streets
890 A, Sullivan BA, Thibaud-Nissen F, Torrance J, Wagner J, Walenz BP, Wenger
891 A, Wood JMD, Xiao C, Yan SM, Young AC, Zarate S, Surti U, McCoy RC,
892 Dennis MY, Alexandrov IA, Gerton JL, O'Neill RJ, Timp W, Zook JM, Schatz
893 MC, Eichler EE, Miga KH, Phillippy AM. 2022. The complete sequence of a
894 human genome. *Science* **376**:44–53.
895 O'Geen H, Squazzo SL, Iyengar S, Blahnik K, Rinn JL, Chang HY, Green R, Farnham
896 PJ. 2007. Genome-wide analysis of KAP1 binding suggests autoregulation of
897 KRAB-ZNFs. *PLoS Genet* **3**:e89.
898 Olson MV. 1999. When less is more: gene loss as an engine of evolutionary change. *Am*
899 *J Hum Genet* **64**:18–23.
900 Pál C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. *Nat Rev*
901 *Genet* **7**:337–348.
902 Paradis E, Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and
903 evolutionary analyses in R. *Bioinformatics* **35**:526–528.
904 Rangasamy D. 2013. Distinctive patterns of epigenetic marks are associated with
905 promoter regions of mouse LINE-1 and LTR retrotransposons. *Mob DNA* **4**:27.
906 Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT,
907 Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL. 2014. A 3D map of the
908 human genome at kilobase resolution reveals principles of chromatin looping.
909 *Cell* **159**:1665–1680.
910 Rice AM, McLysaght A. 2017. Dosage sensitivity is a major determinant of human
911 copy number variant pathogenicity. *Nat Commun* **8**:14366.
912 Roux J, Liu J, Robinson-Rechavi M. 2017. Selective Constraints on Coding Sequences
913 of Nervous System Genes Are a Major Determinant of Duplicate Gene
914 Retention in Vertebrates. *Mol Biol Evol* **34**:2773–2791.
915 Schaibley VM, Zawistowski M, Wegmann D, Ehm MG, Nelson MR, St Jean PL,
916 Abecasis GR, Novembre J, Zöllner S, Li JZ. 2013. The influence of genomic
917 context on mutation patterns in the human genome inferred from rare variants.
918 *Genome Res* **23**:1974–1984.
919 Schield DR, Card DC, Hales NR, Perry BW, Pasquesi GM, Blackmon H, Adams RH,
920 Corbin AB, Smith CF, Ramesh B, Demuth JP, Betrán E, Tollis M, Meik JM,
921 Mackessy SP, Castoe TA. 2019. The origins and evolution of chromosomes,

dosage compensation, and mechanisms underlying venom regulation in snakes. *Genome Res* **29**:590–601.

Seplyarskiy VB, Sunyaev S. 2021. The origin of human mutation in light of genomic data. *Nat Rev Genet* **22**:672–686.

Sharma V, Hecker N, Roscito JG, Foerster L, Langer BE, Hiller M. 2018. A genomics approach reveals insights into the importance of gene losses for mammalian adaptations. *Nat Commun* **9**:1215.

Shen X-X, Opulente DA, Kominek J, Zhou X, Steenwyk JL, Buh KV, Haase MAB, Wisecaver JH, Wang M, Doering DT, Boudouris JT, Schneider RM, Langdon QK, Ohkuma M, Endoh R, Takashima M, Manabe R-I, Čadež N, Libkind D, Rosa CA, DeVirgilio J, Hulfachor AB, Groenewald M, Kurtzman CP, Hittinger CT, Rokas A. 2018. Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum. *Cell* **175**:1533-1545.e20.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**:308–311.

Simakov O, Marlétaz F, Yue J-X, O'Connell B, Jenkins J, Brandt A, Calef R, Tung C-H, Huang T-K, Schmutz J, Satoh N, Yu J-K, Putnam NH, Green RE, Rokhsar DS. 2020. Deeply conserved synteny resolves early events in vertebrate evolution. *Nat Ecol Evol* **4**:820–830.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**:3210–3212.

Slotkin RK, Martienssen R. 2007. Transposable elements and the epigenetic regulation of the genome. *Nat Rev Genet* **8**:272–285.

Stamatoyannopoulos JA, Adzhubei I, Thurman RE, Kryukov GV, Mirkin SM, Sunyaev SR. 2009. Human mutation rate associated with DNA replication timing. *Nat Genet* **41**:393–395.

Steinegger M, Söding J. 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* **35**:1026–1028.

Terekhanova NV, Seplyarskiy VB, Soldatov RA, Bazykin GA. 2017. Evolution of Local Mutation Rate and Its Determinants. *Mol Biol Evol* **34**:1100–1109.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**:562–578.

Underwood CJ, Henderson IR, Martienssen RA. 2017. Genetic and epigenetic variation of transposable elements in Arabidopsis. *Curr Opin Plant Biol* **36**:135–141.

van Schaik T, Vos M, Peric-Hupkes D, Hn Celie P, van Steensel B. 2020. Cell cycle dynamics of lamina-associated DNA. *EMBO Rep* **21**:e50636.

van Steensel B, Belmont AS. 2017. Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression. *Cell* **169**:780–791.

Vogel MJ, Guelen L, de Wit E, Peric-Hupkes D, Lodén M, Talhout W, Feenstra M, Abbas B, Classen A-K, van Steensel B. 2006. Human heterochromatin proteins form large domains containing KRAB-ZNF genes. *Genome Res* **16**:1493–1504.

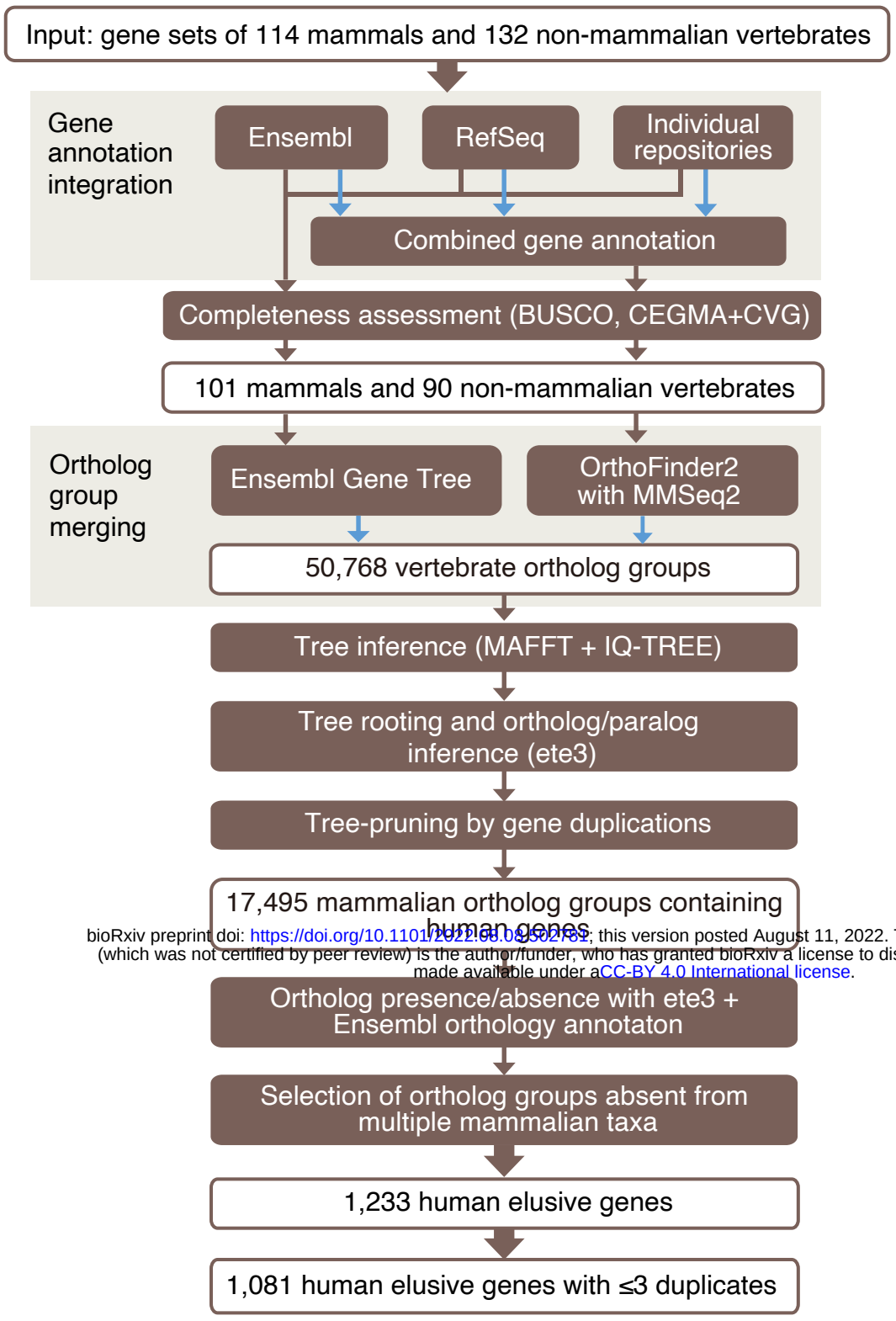
Yang J, Gu Z, Li W-H. 2003. Rate of protein evolution versus fitness effect of gene deletion. *Mol Biol Evol* **20**:772–774.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**:1586–1591.

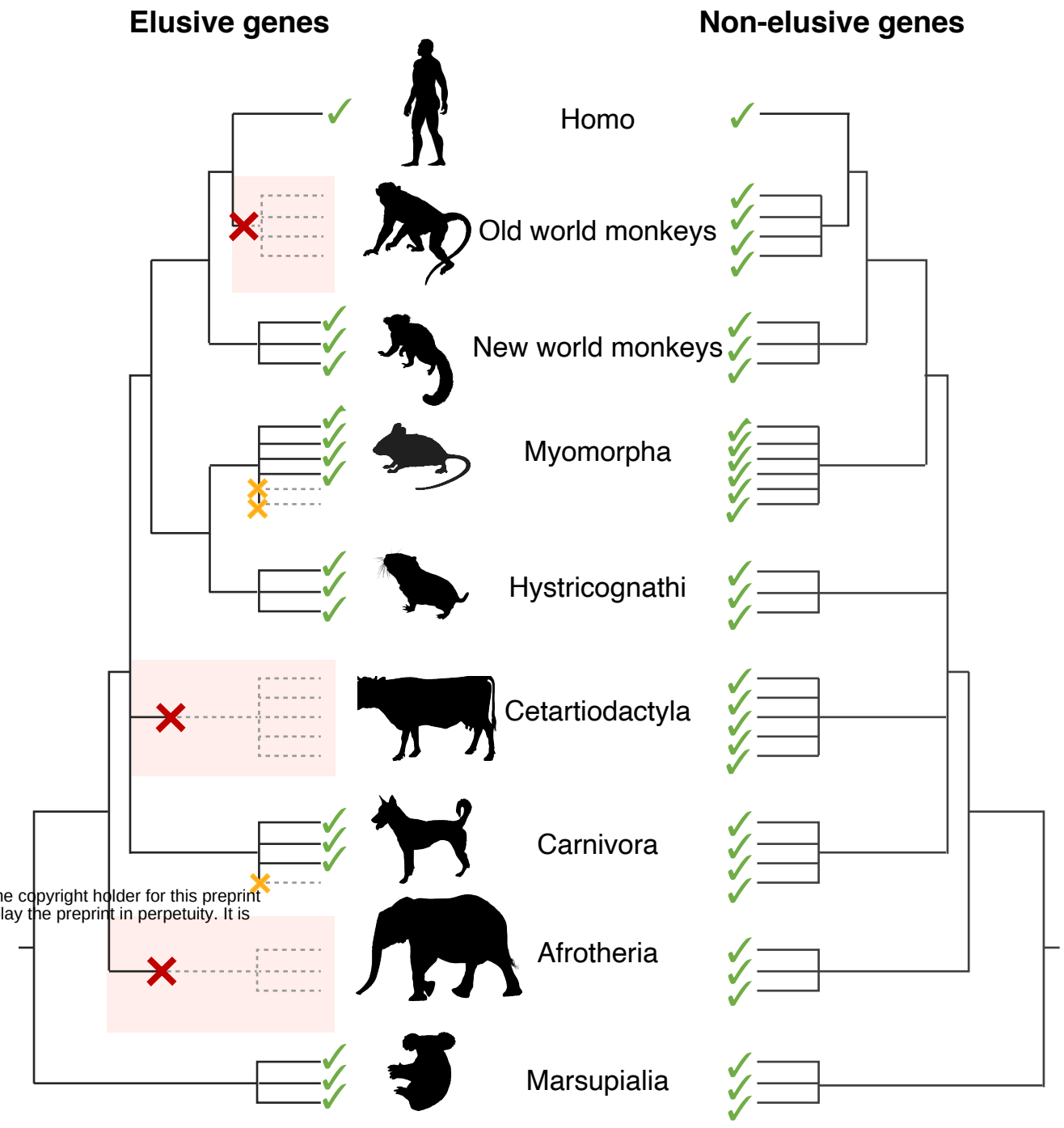
Yoshihara M, Jiang L, Akatsuka S, Suyama M, Toyokuni S. 2014. Genome-wide profiling of 8-oxoguanine reveals its association with spatial positioning in nucleus. *DNA Res* **21**:603–612.

Zheng X, Hu J, Yue S, Kristiani L, Kim M, Sauria M, Taylor J, Kim Y, Zheng Y. 2018. Lamins Organize the Global Three-Dimensional Genome from the Nuclear Periphery. *Mol Cell* **71**:802-815.e7.

a



b



c

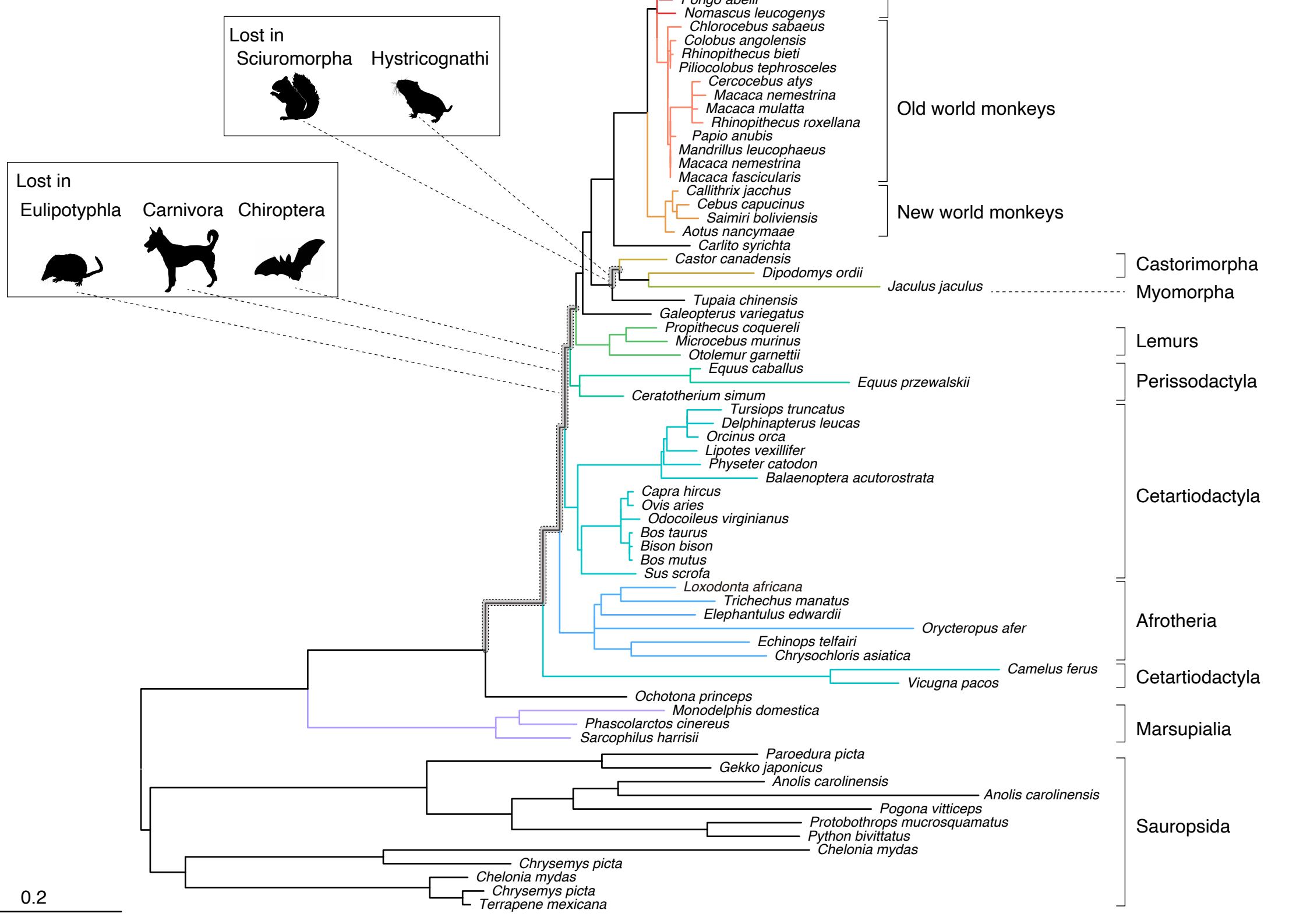


Figure 1

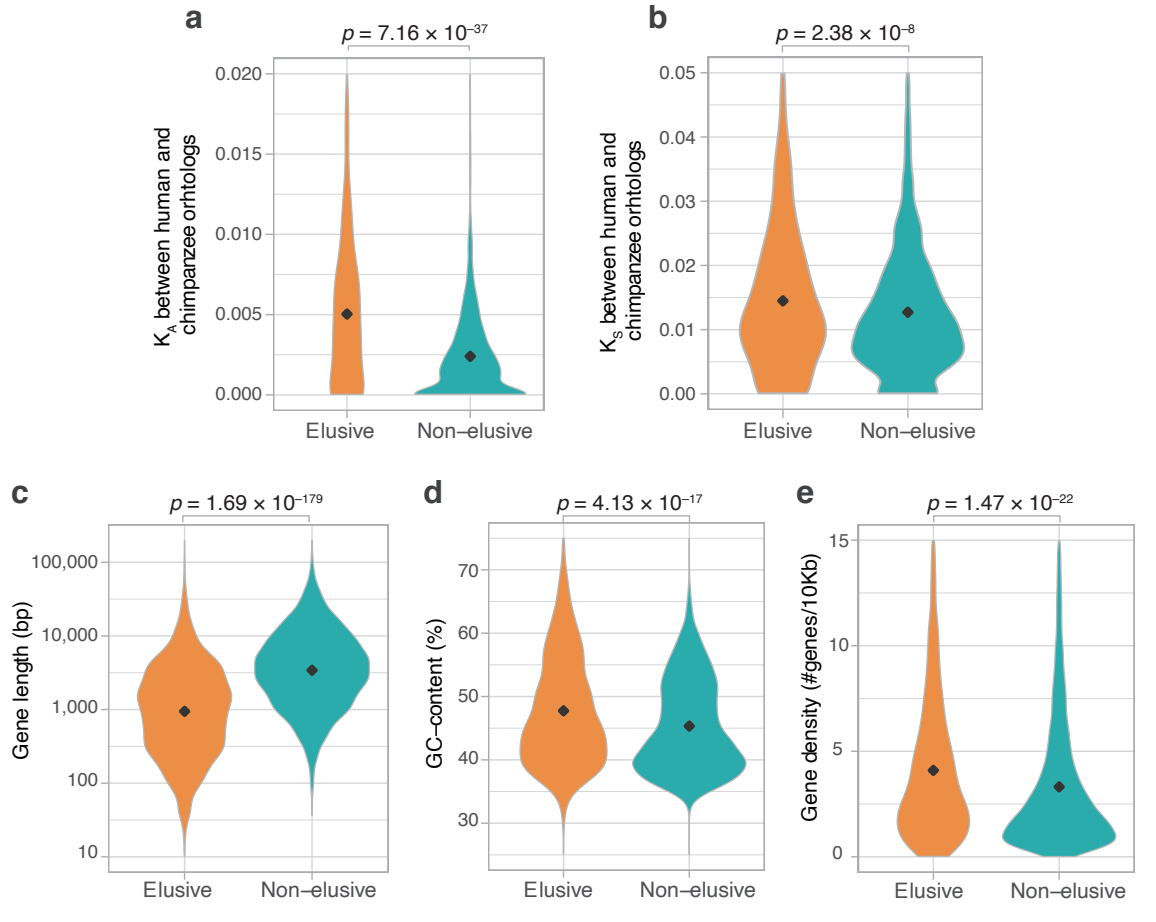

a
p = 7.16 × 10−37
K_A between human and chimpanzee orhtologs
0.020
0.015
0.010
0.005
0.000
Elusive
Non−elusive
b
p = 2.38 × 10−8
K_S between human and chimpanzee orhtologs
0.05
0.04
0.03
0.02
0.01
0.00
Elusive
Non−elusive
c
p = 1.69 × 10−179
Gene length (bp)
100,000
10,000
1,000
100
10
Elusive
Non−elusive
d
p = 4.13 × 10−17
GC−content (%)
70
60
50
40
30
Elusive
Non−elusive
e
p = 1.47 × 10−22
Gene density (#genes/10Kb)
15
10
5
0
Elusive
Non−elusive

Figure 2

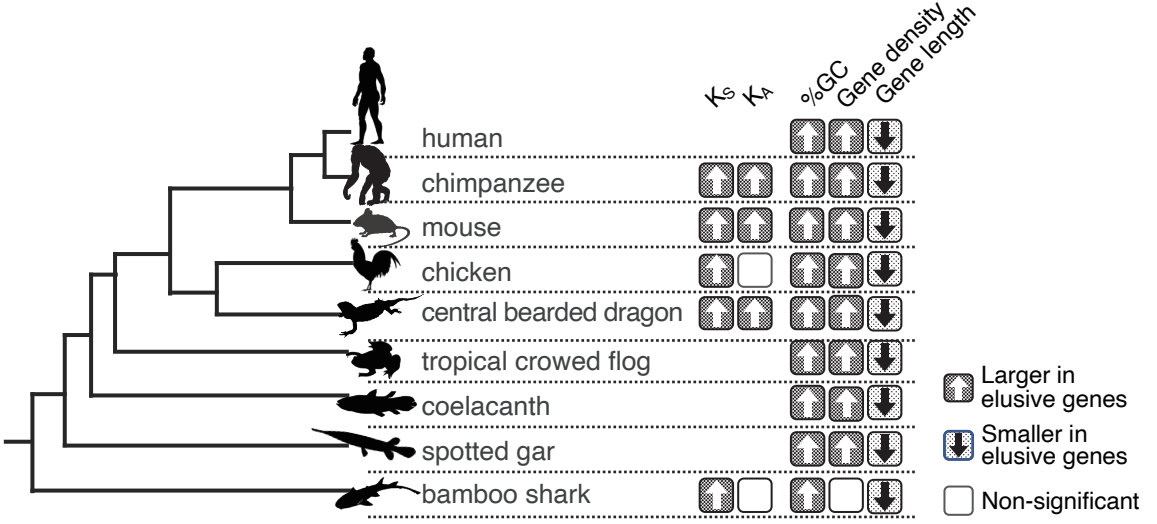

K_S
K_A
%GC
Gene density
Gene length
human
chimpanzee
mouse
chicken
central bearded dragon
tropical crowed flog
coelacanth
spotted gar
bamboo shark
Larger in elusive genes
Smaller in elusive genes
Non-significant

Figure 3

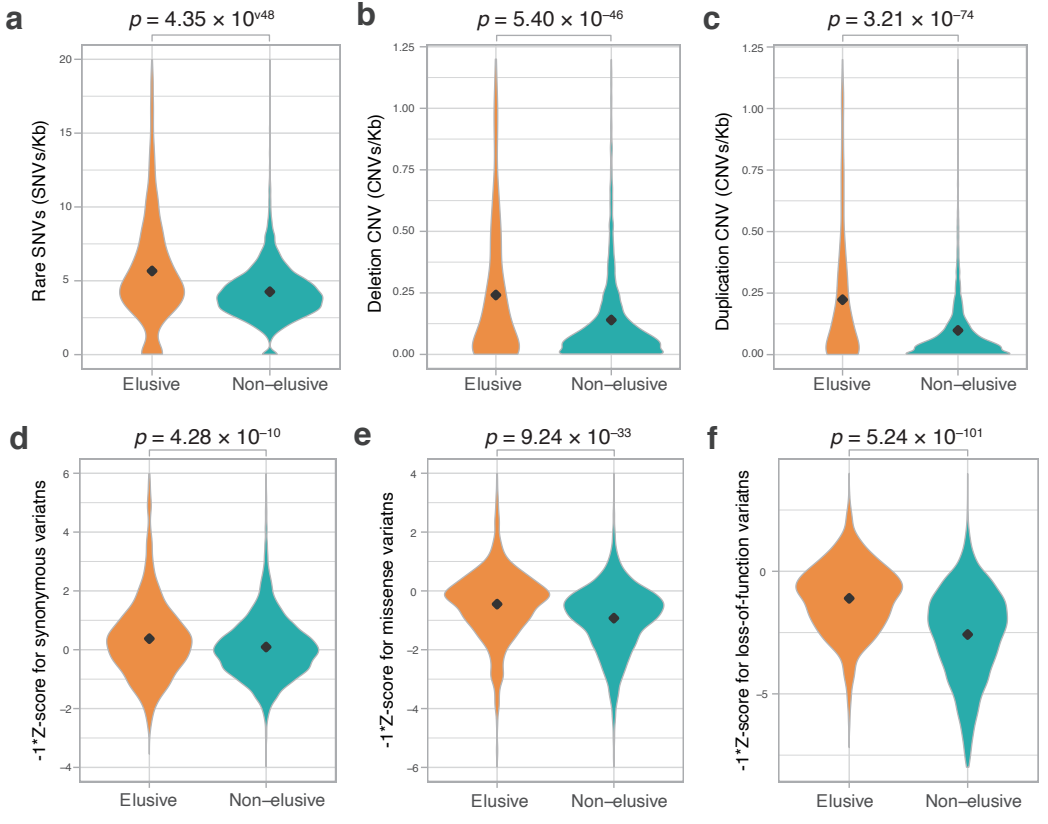

a
p = 4.35 × 10v48
Rare SNVs (SNVs/Kb)
20
15
10
5
0
Elusive
Non−elusive
b
p = 5.40 × 10−46
Deletion CNV (CNVs/Kb)
1.25
1.00
0.75
0.50
0.25
0.00
Elusive
Non−elusive
c
p = 3.21 × 10−74
Duplication CNV (CNVs/Kb)
1.25
1.00
0.75
0.50
0.25
0.00
Elusive
Non−elusive
d
p = 4.28 × 10−10
-1*Z-score for synonymous variatns
6
4
2
0
−2
−4
Elusive
Non−elusive
e
p = 9.24 × 10−33
-1*Z-score for missense variatns
4
2
0
−2
−4
−6
Elusive
Non−elusive
f
p = 5.24 × 10−101
-1*Z-score for loss-of-function variatns
0
−5
Elusive
Non−elusive

Figure 4

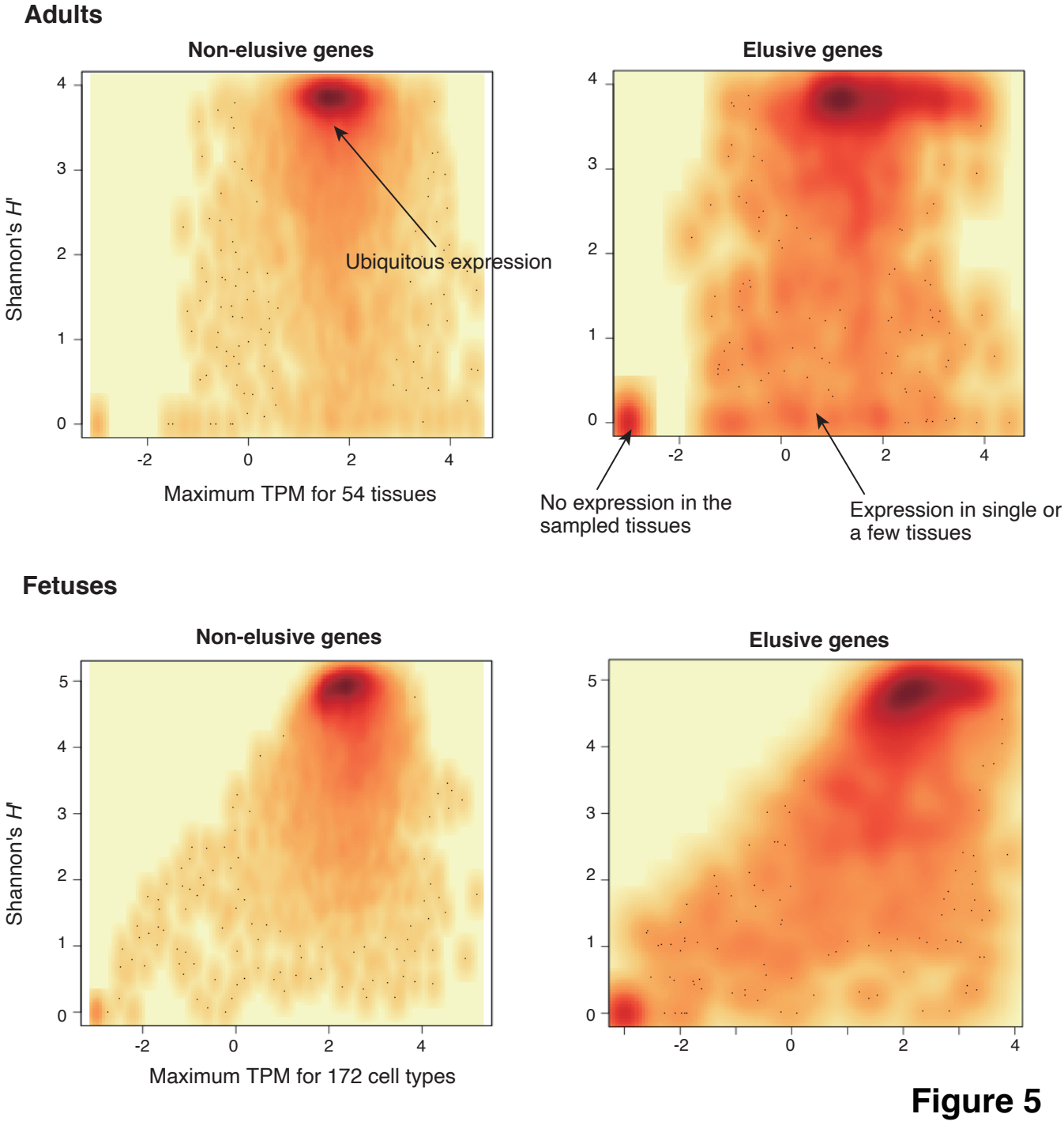

Adults
Non-elusive genes
Elusive genes
Shannon's H'
Maximum TPM for 54 tissues
Ubiquitous expression
No expression in the sampled tissues
Expression in single or a few tissues
Fetuses
Non-elusive genes
Elusive genes
Shannon's H'
Maximum TPM for 172 cell types

Figure 5

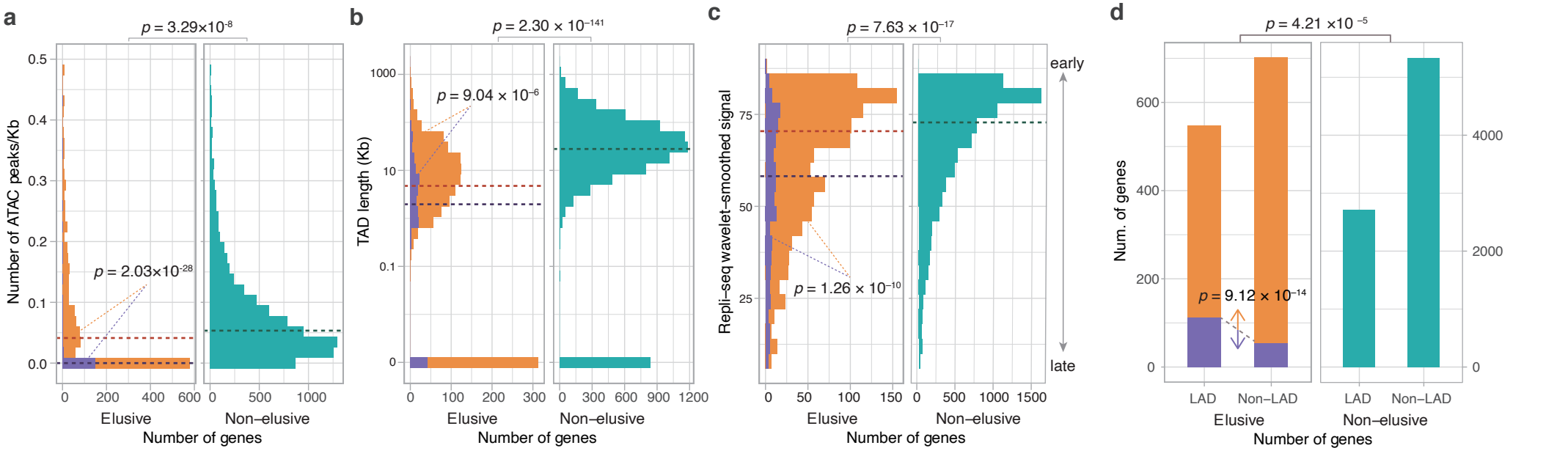

a
p = 3.29×10-8
Number of ATAC peaks/Kb
p = 2.03×10-28
Elusive
Non−elusive
Number of genes
b
p = 2.30 × 10−141
TAD length (Kb)
p = 9.04 × 10−6
Elusive
Non−elusive
Number of genes
c
p = 7.63 × 10−17
Repli−seq wavelet−smoothed signal
early
late
p = 1.26 × 10−10
Elusive
Non−elusive
Number of genes
d
p = 4.21 ×10 −5
Num. of genes
p = 9.12 × 10−14
LAD
Non−LAD
LAD
Non−LAD
Elusive
Non−elusive
Number of genes

Figure 6

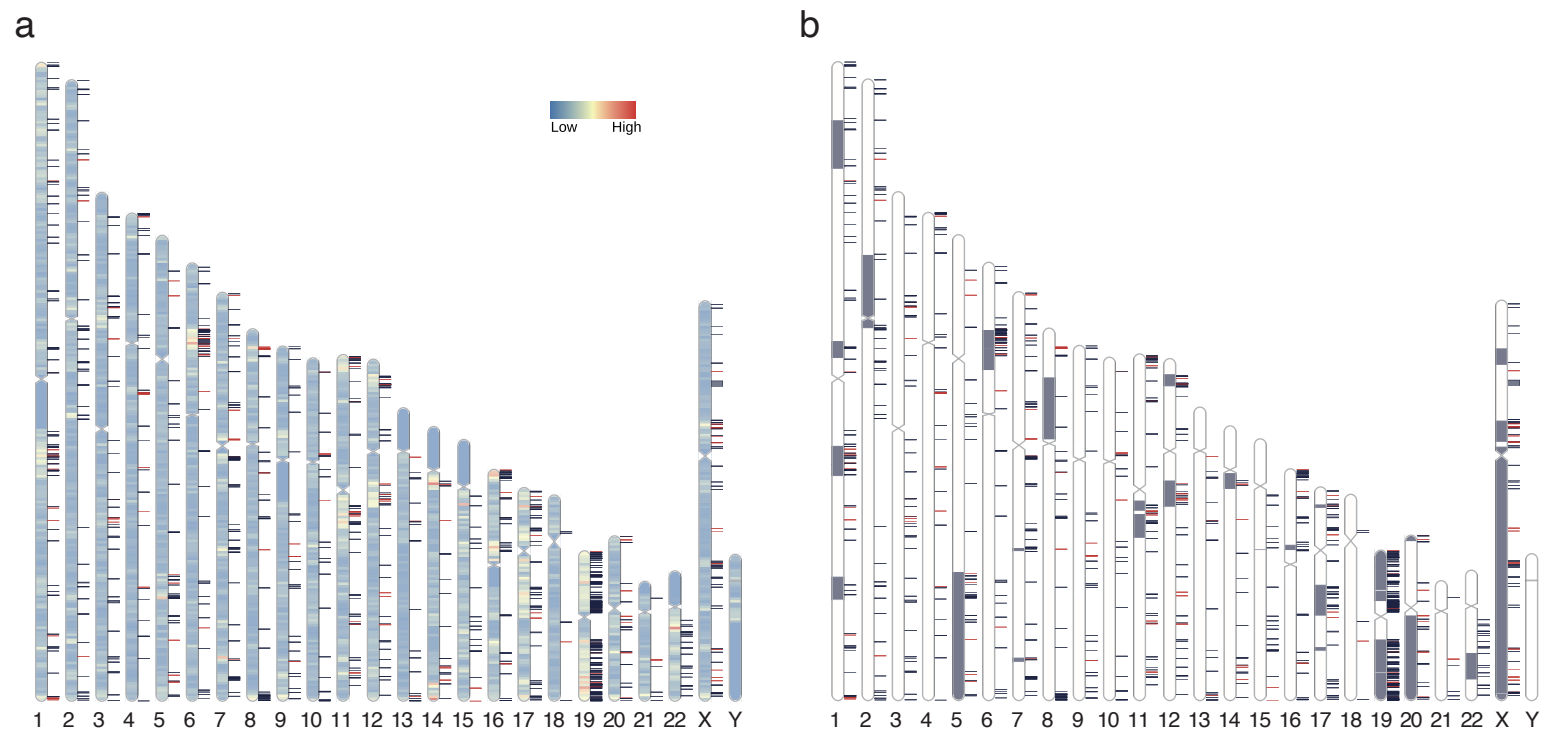

a
Low
High
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
b
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y

Figure 7

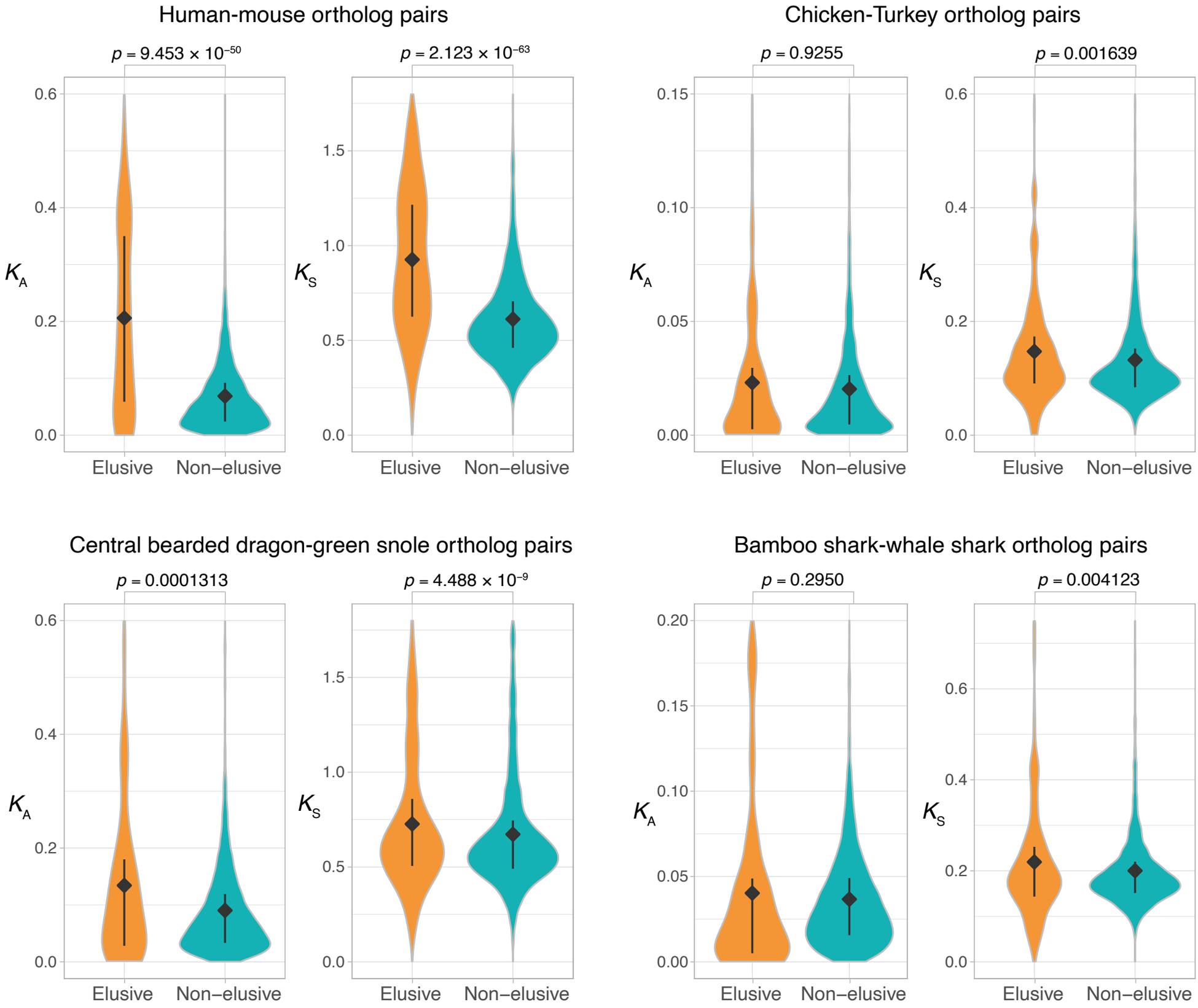
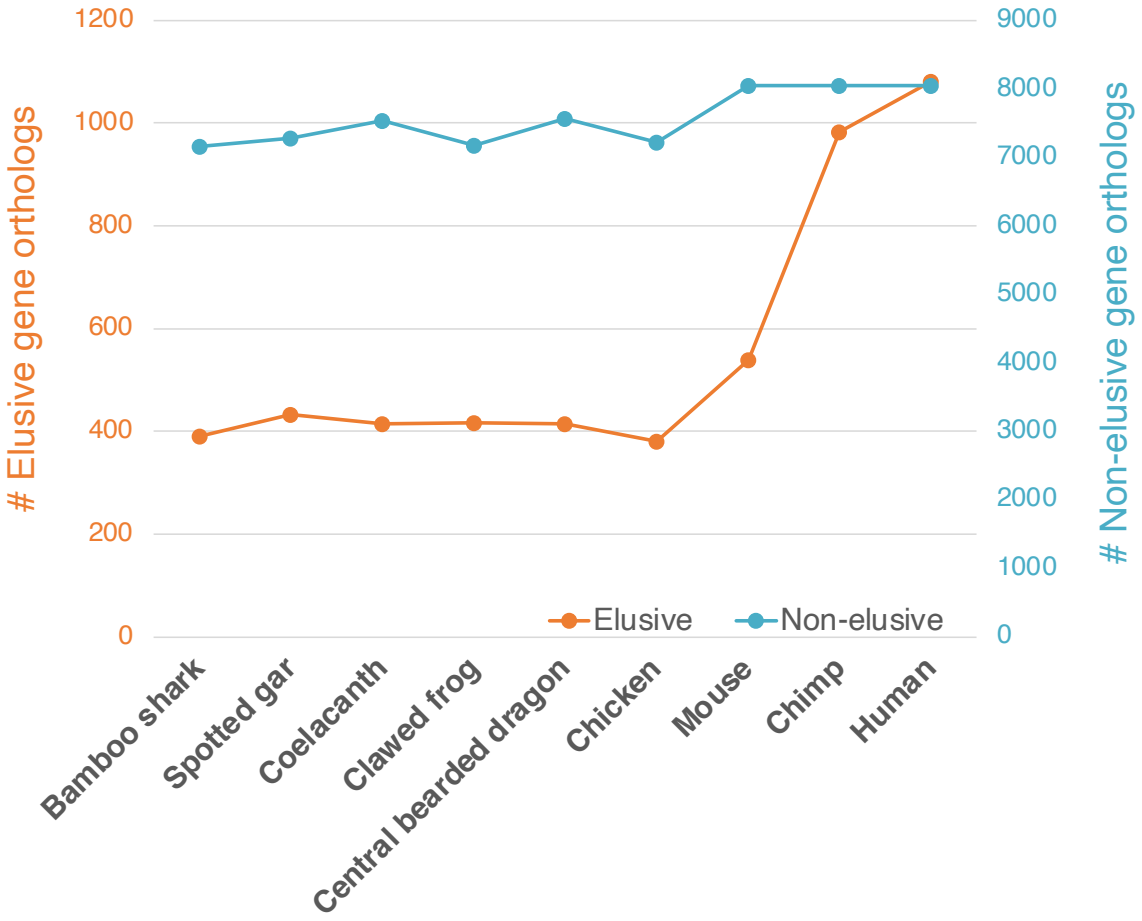


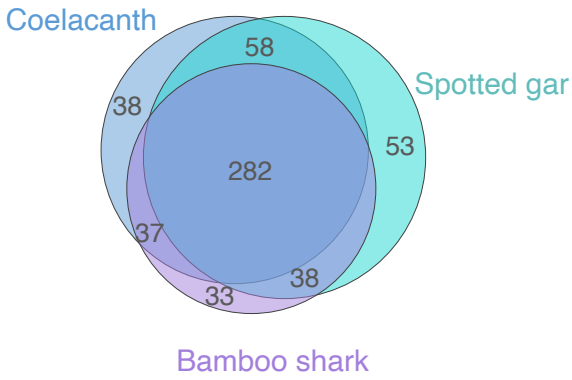
Figure 2–figure supplement 1

a

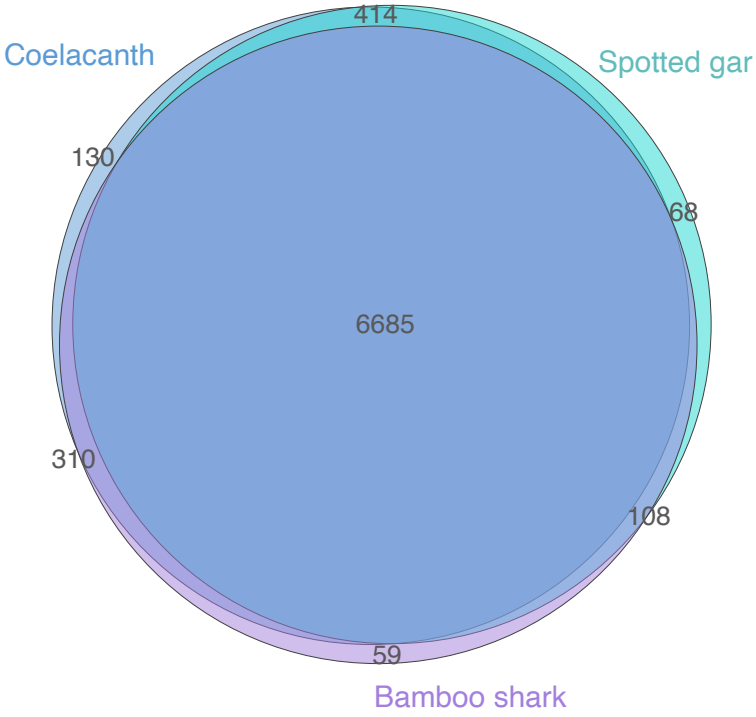


b

Elusive gene orthologs

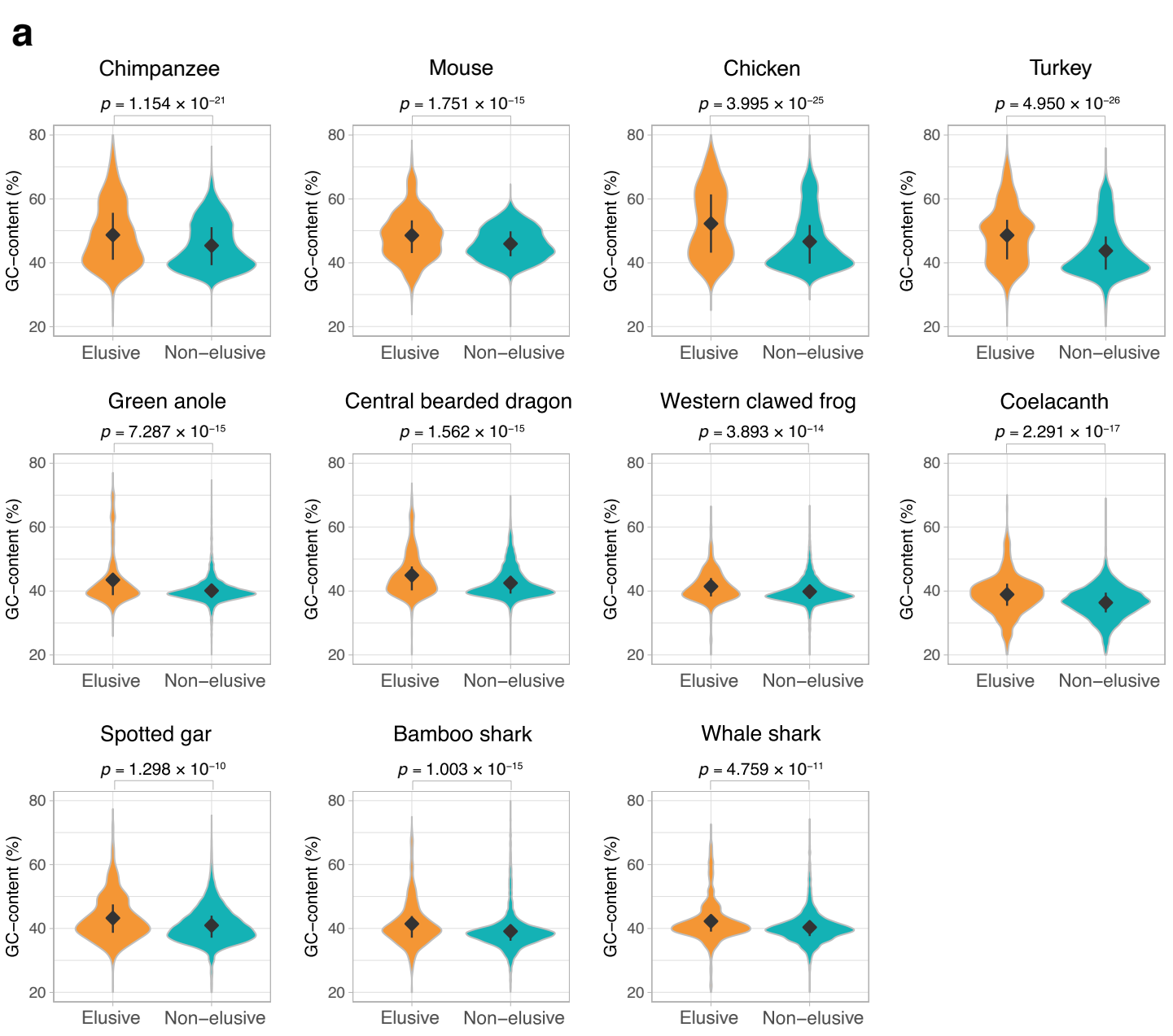


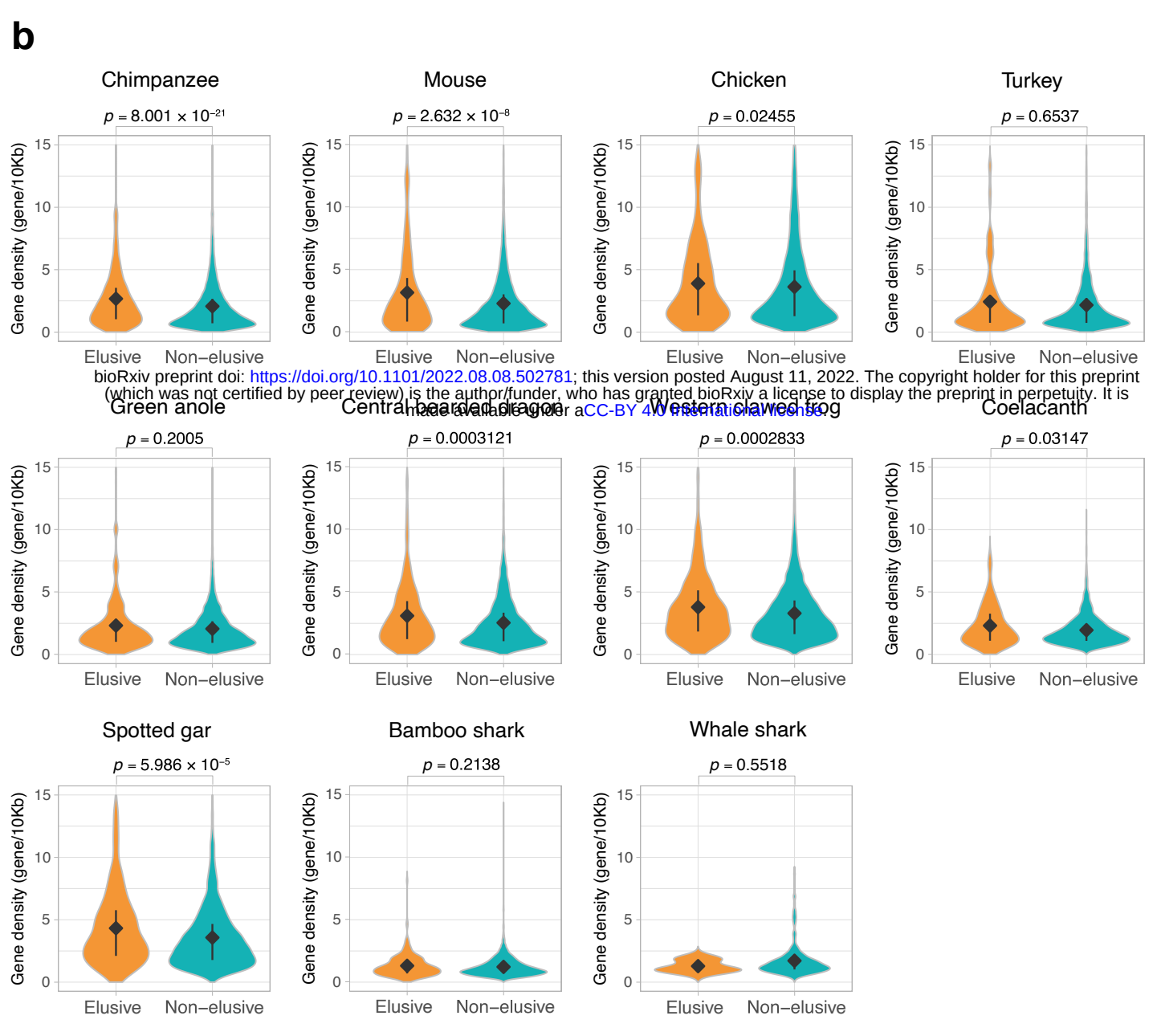
Non-elusive gene orthologs



	3 species	1 or 2 species
Elusive genes	282	257
Non-elusive genes	6685	1089

Figure 3–figure supplement 1





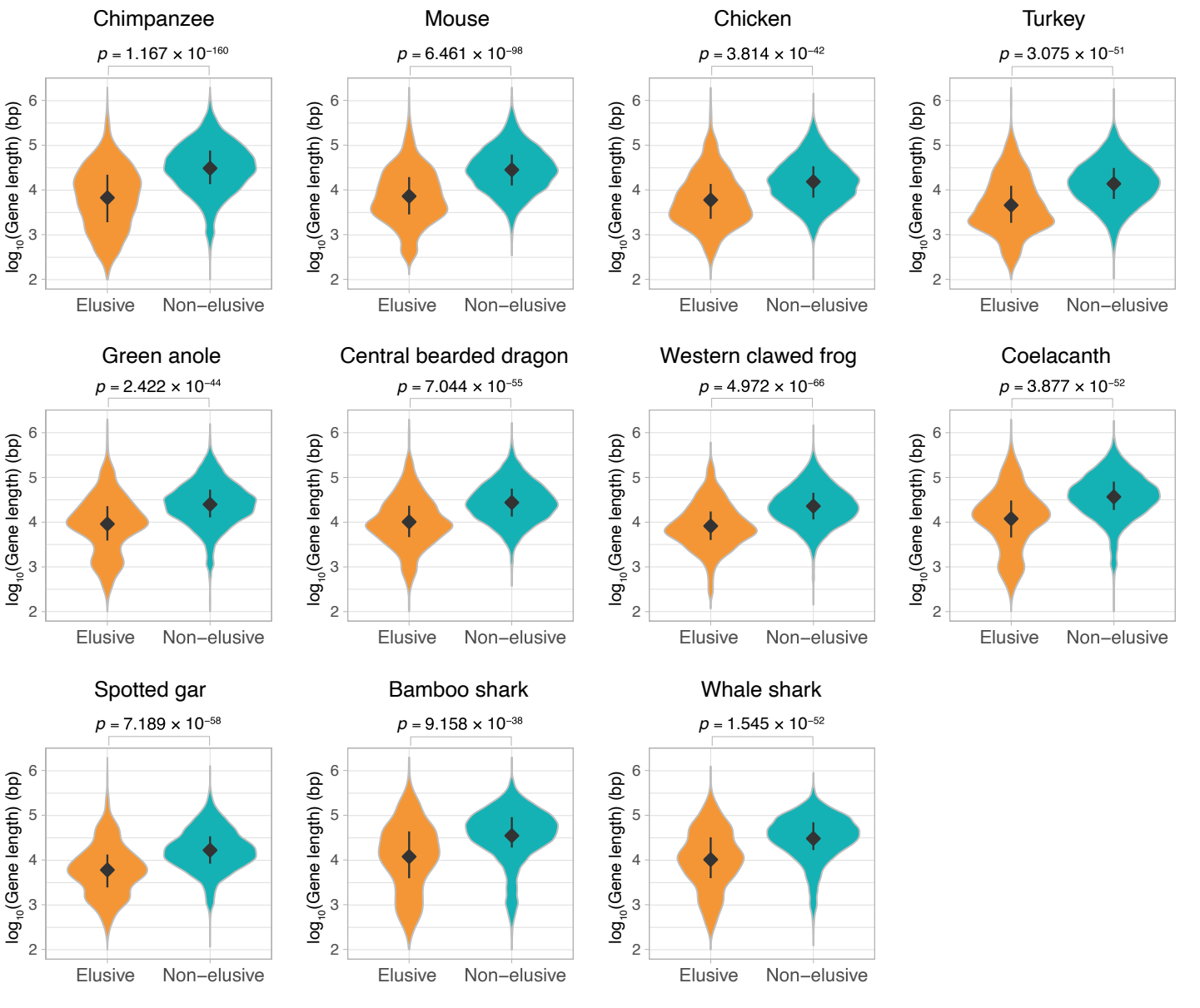


Figure 3–figure supplement 2

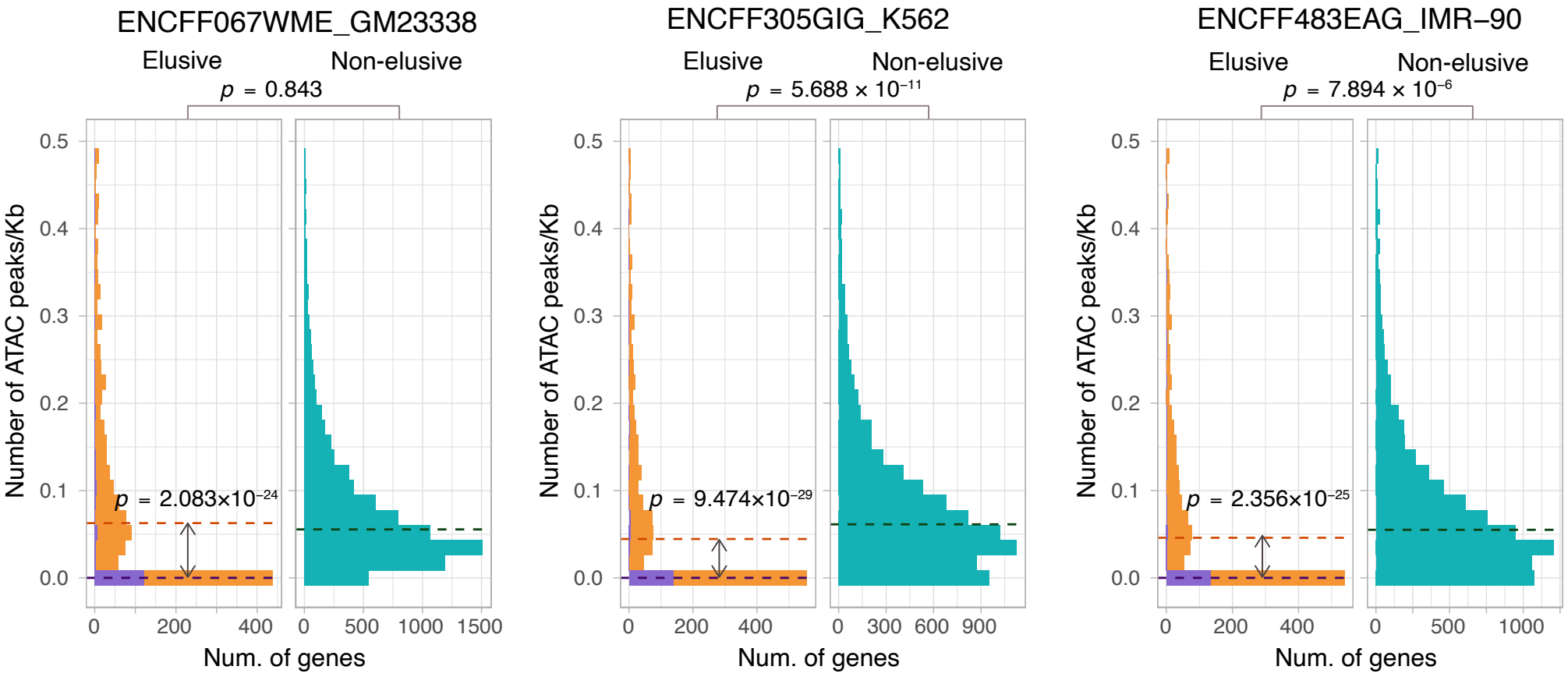
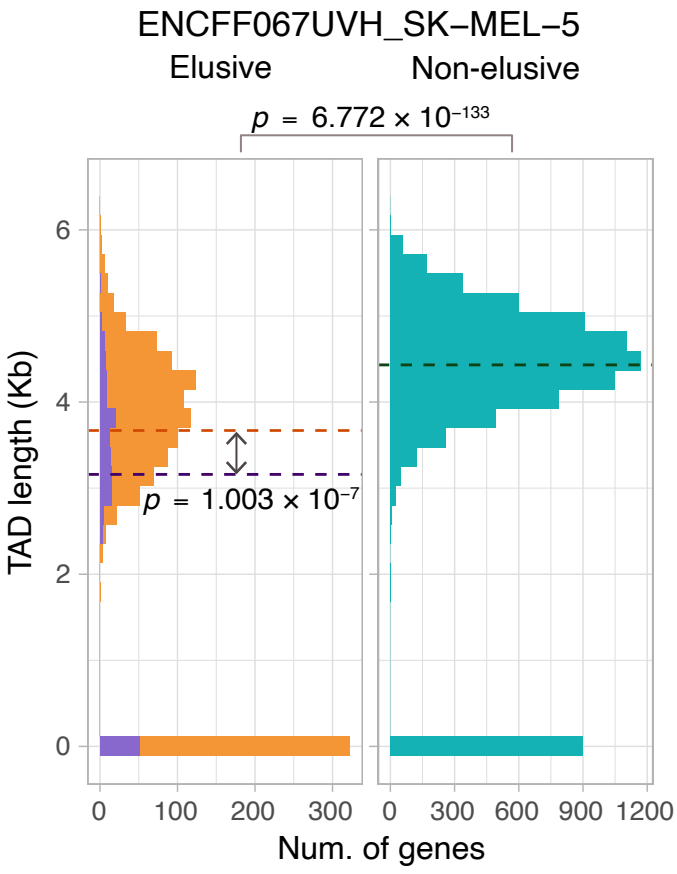
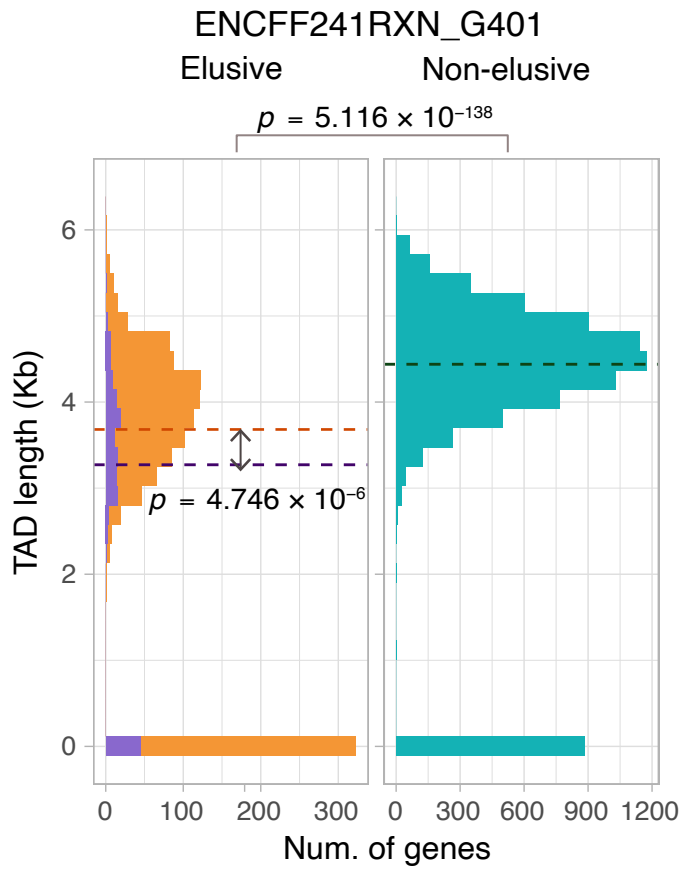
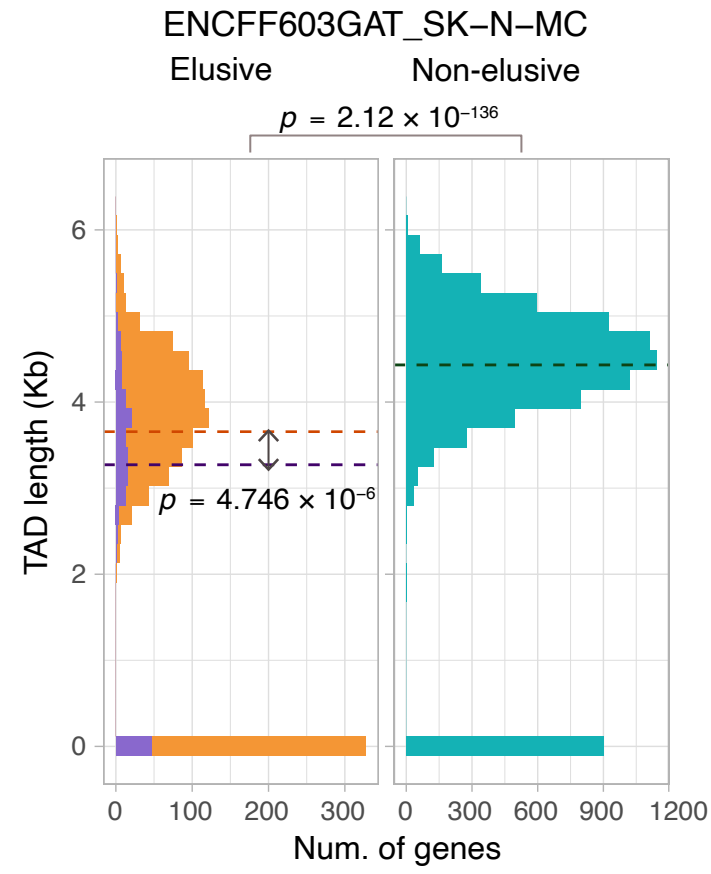
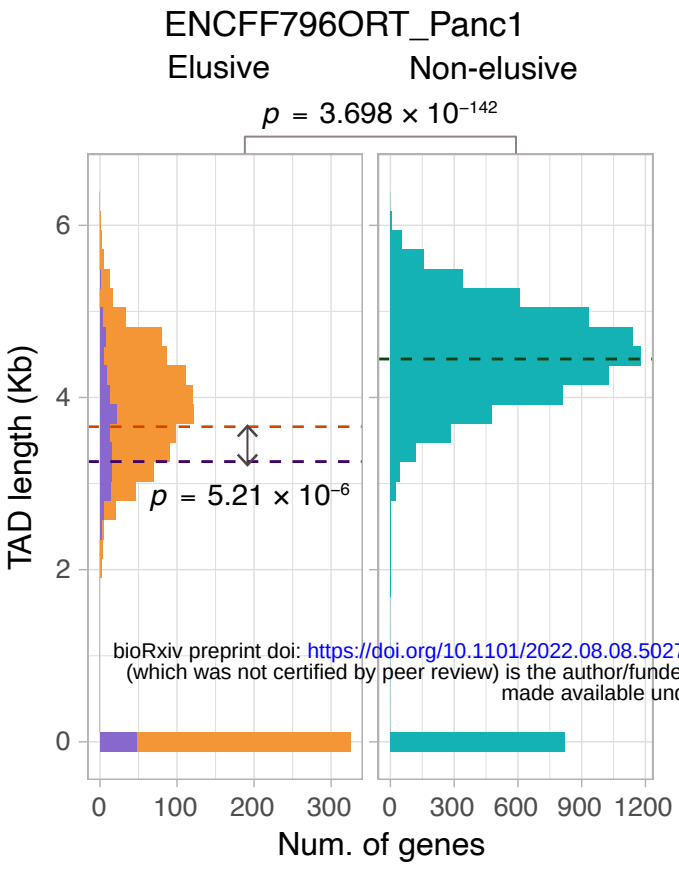


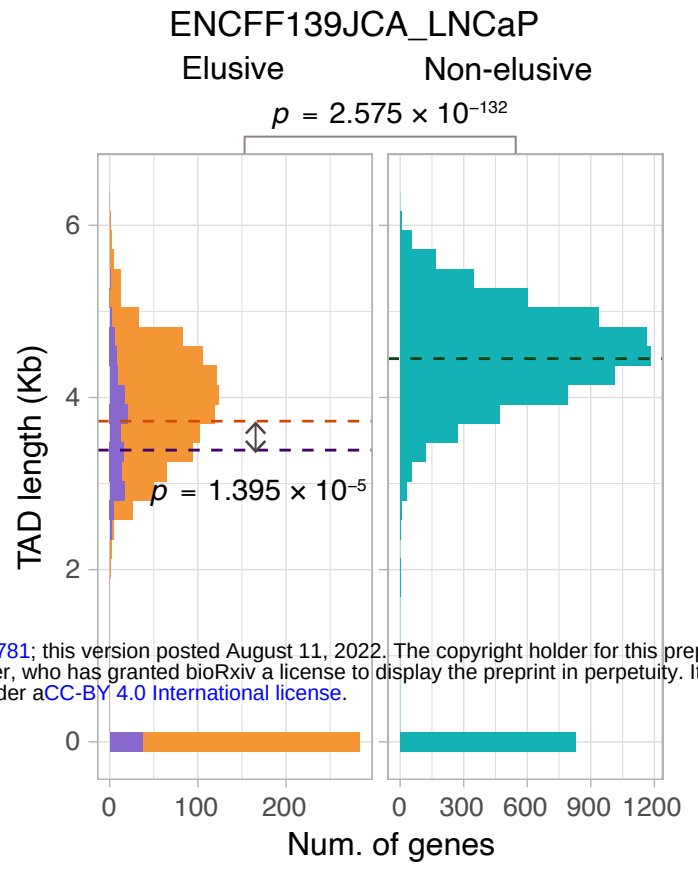
Figure 6–figure supplement 1

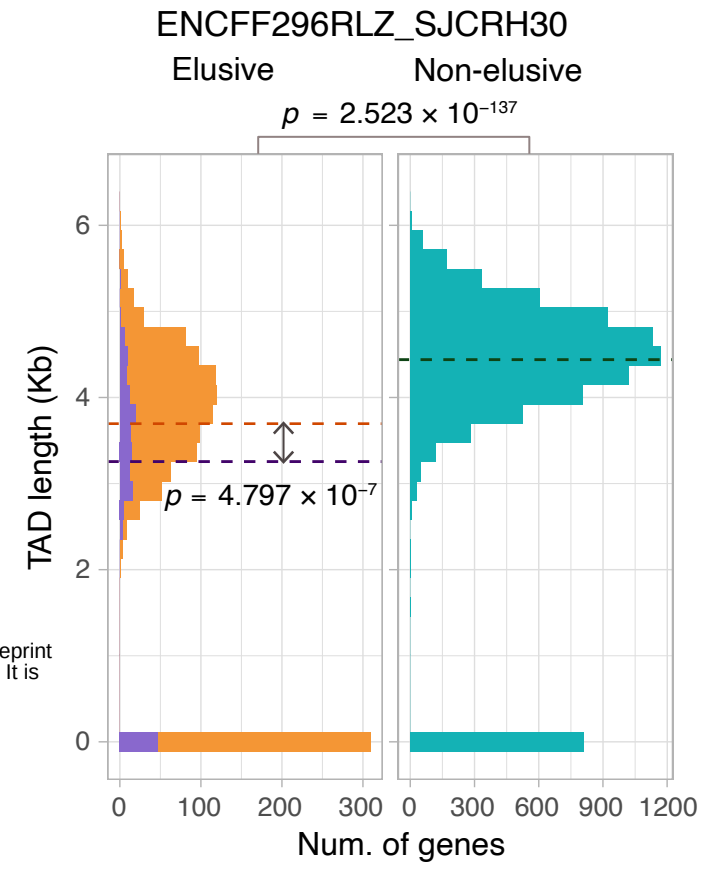


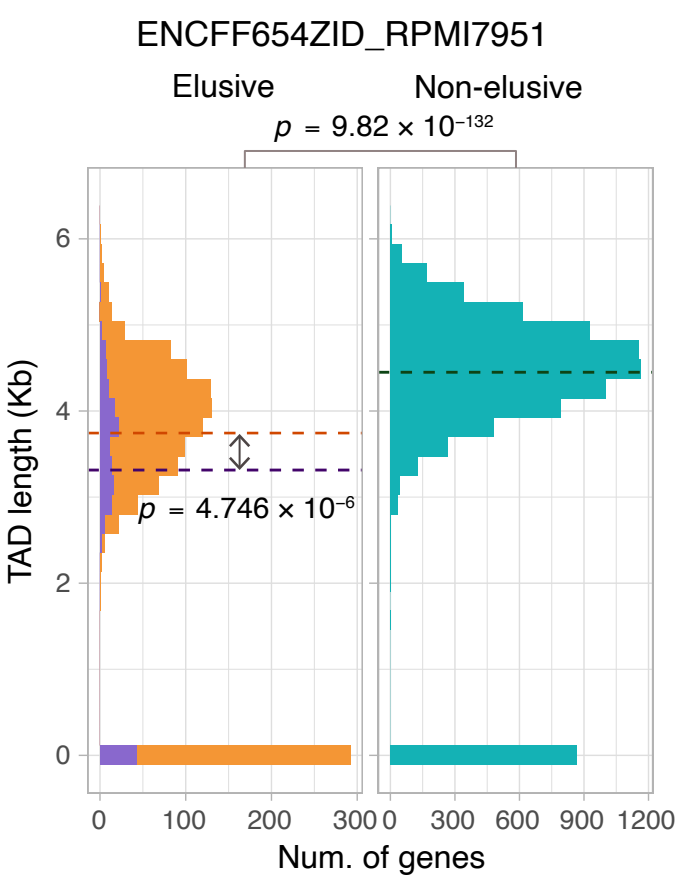


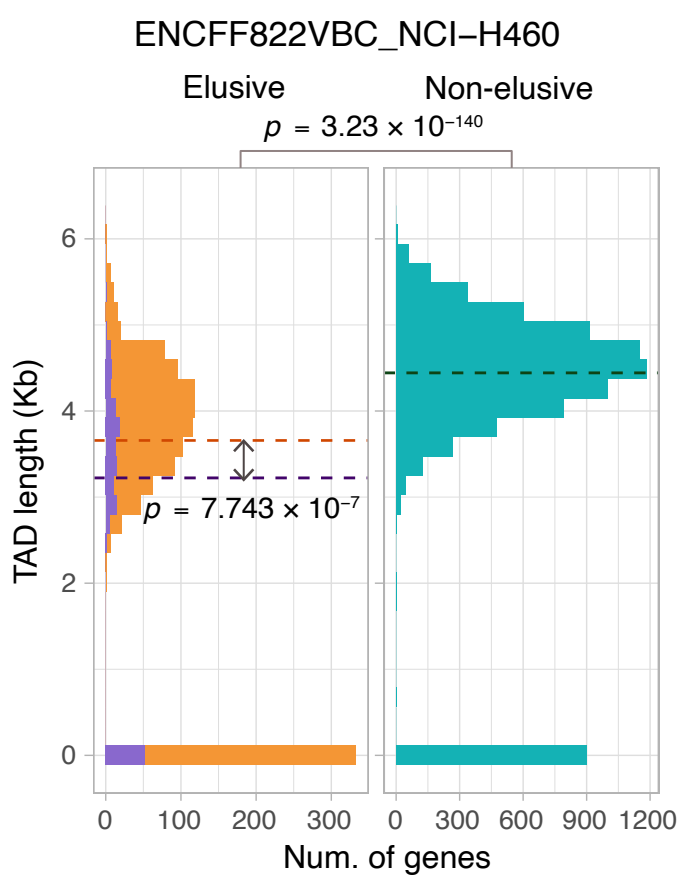


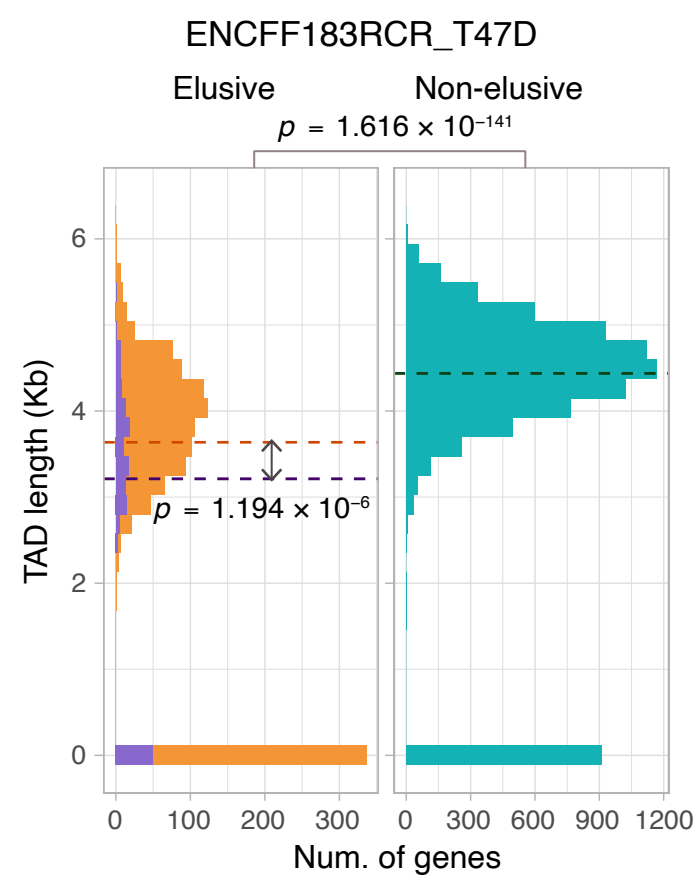












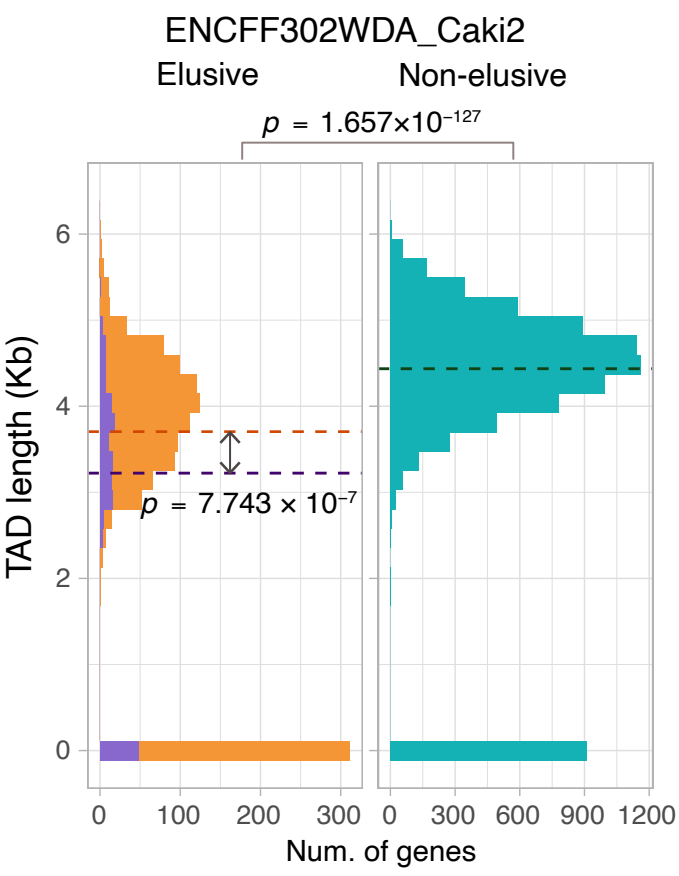


Figure 6–figure supplement 2

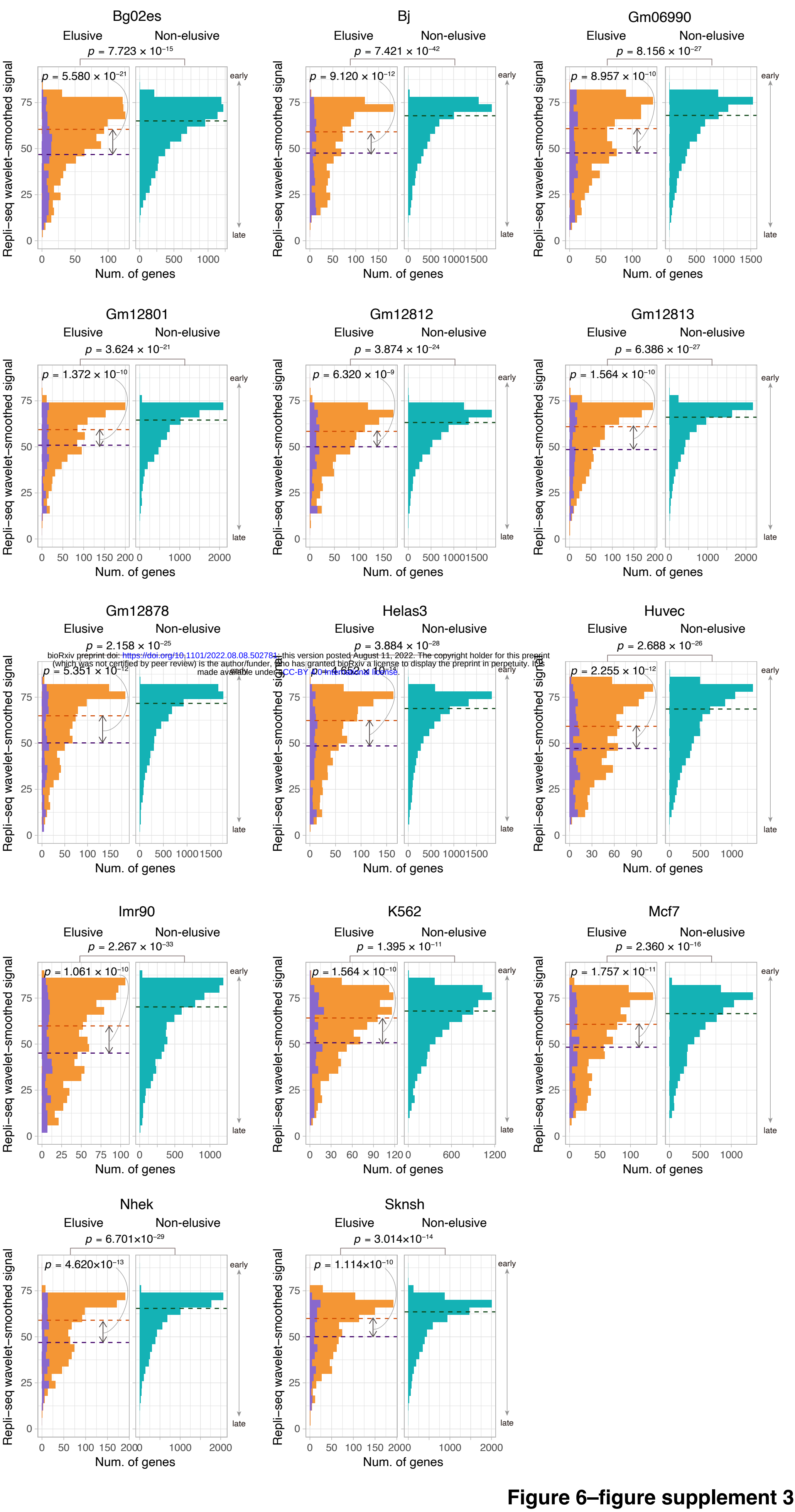
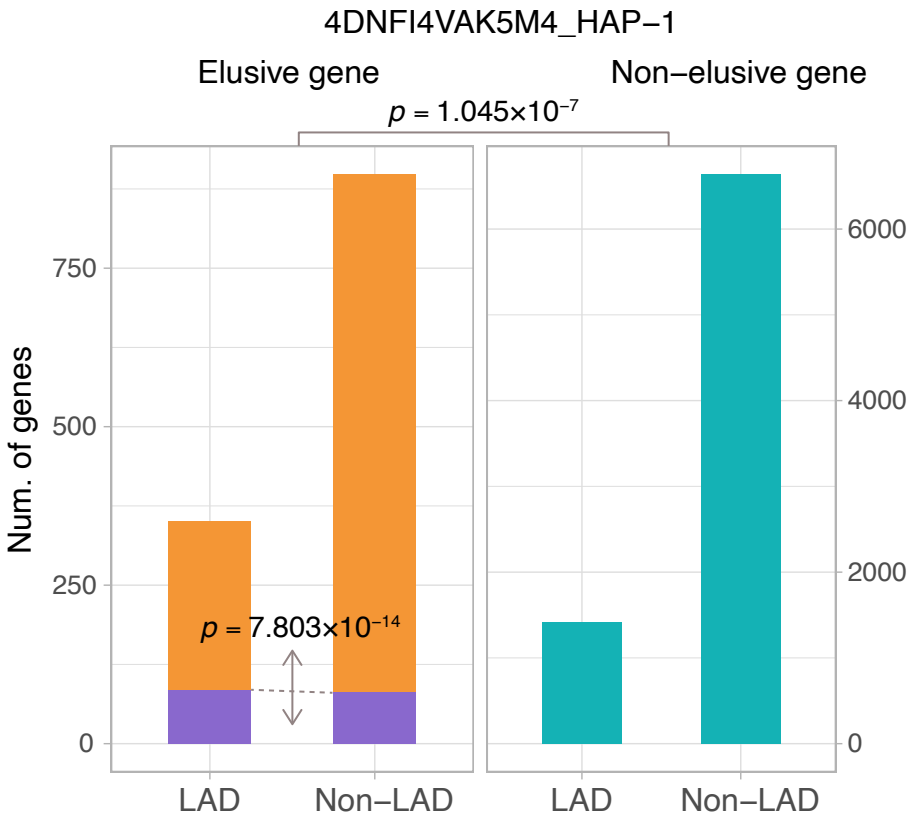
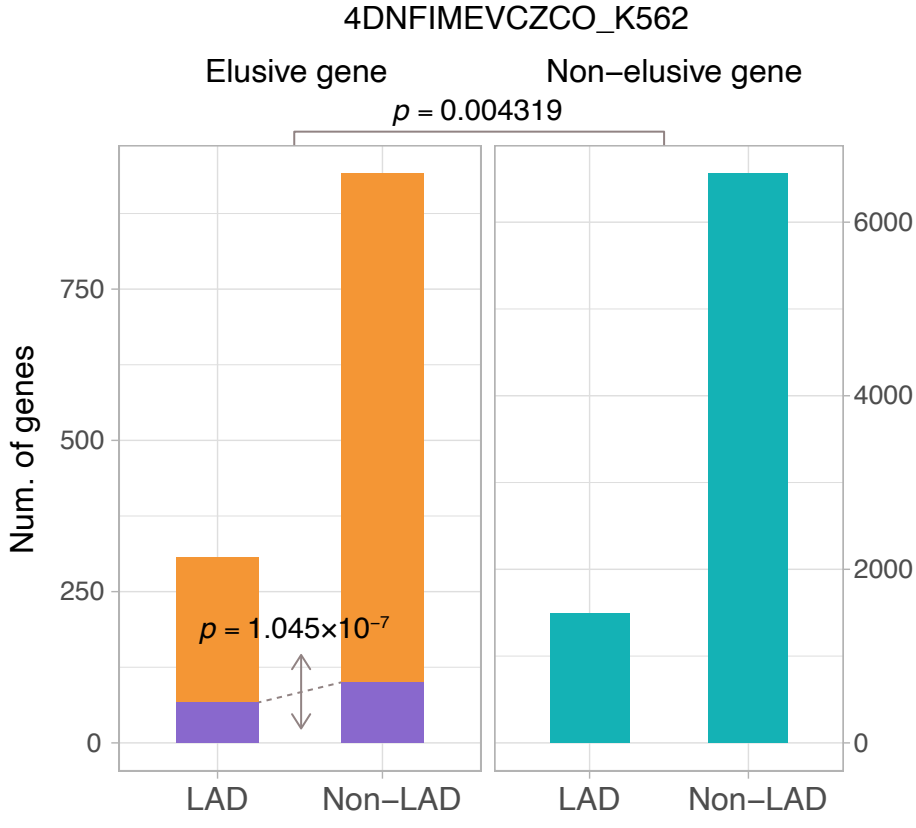


Figure 6–figure supplement 3





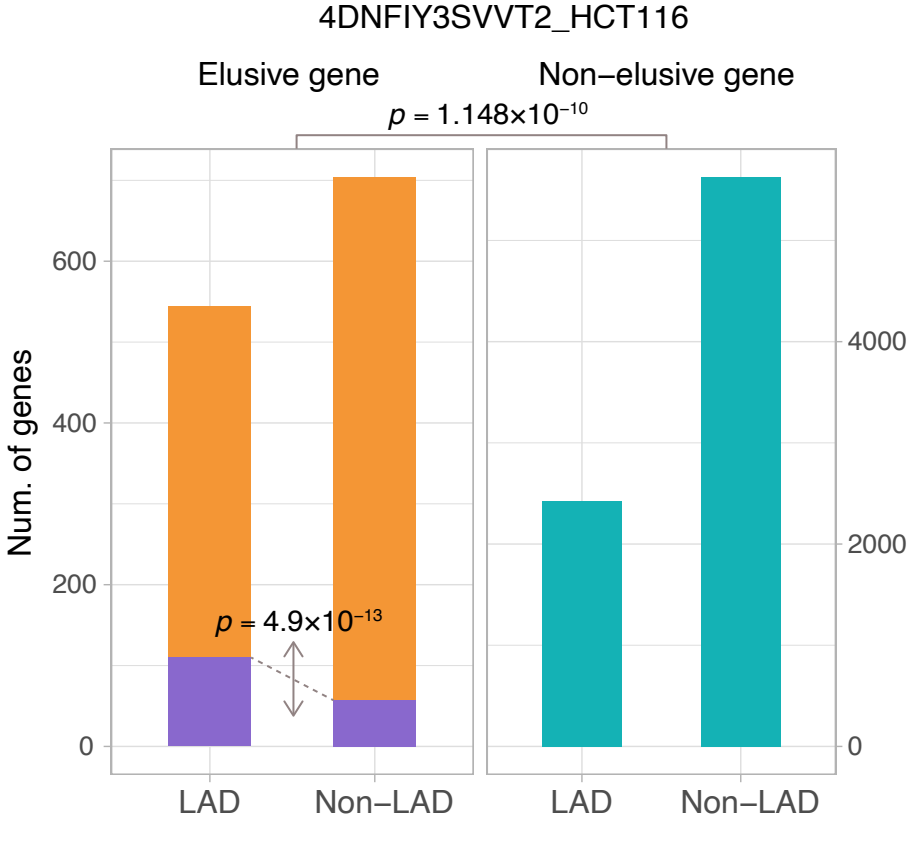


Figure 6–figure supplement 4

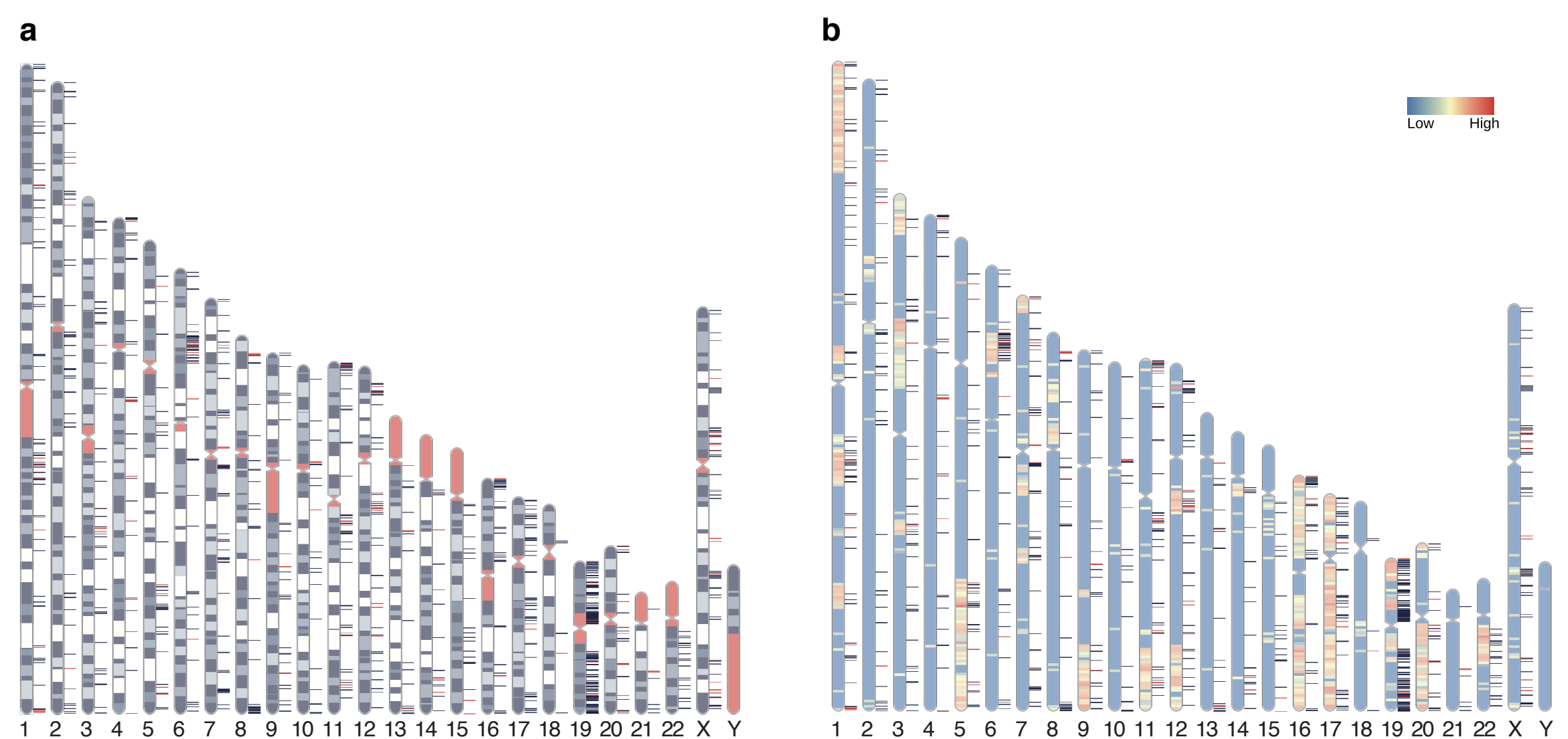


Figure 7–figure supplement 1