Enhancers facilitate the birth of *de novo* genes and their functional integration into regulatory networks

Paco Majic[1,2] & Joshua L. Payne[1,2,*]

1. Institute of Integrative Biology, ETH Zurich, Switzerland

2. Swiss Institute of Bioinformatics, Lausanne, Switzerland

* Correspondence to: joshua.payne@env.ethz.ch

Abstract

Regulatory networks control the spatiotemporal gene expression patterns that give rise to and define the individual cell types of multicellular organisms. In Eumetazoa, distal regulatory elements called enhancers play a key role in determining the structure of such networks, particularly the wiring diagram of "who regulates whom." Mutations that affect enhancer activity can therefore rewire regulatory networks, potentially causing changes in gene expression that may be adaptive. Here, we use single-cell transcriptomic and chromatin accessibility data from mouse to show that enhancers play an additional role in the evolution of regulatory networks: They facilitate network growth by creating transcriptionally active regions of open chromatin that are conducive to *de novo* gene evolution. Specifically, our comparative transcriptomic analysis with three other mammalian species shows that young, mouse-specific transcribed open reading frames are preferentially located near enhancers, whereas older open reading frames are not. Interactions with enhancers are then gained incrementally over macro-evolutionary timescales, helping integrate new genes into existing regulatory networks. Taken together, our results highlight a dual role of enhancers in expanding and rewiring gene regulatory networks.

Introduction

Enhancers are a defining characteristic of eumetazoan gene regulatory networks. They recruit transcription factors and cofactors that "loop out" DNA to bind core promoters and increase the expression of target genes [1, 2], thus mediating interactions between genes. Such interactions are highly dynamic throughout development, facilitating the differential deployment of distinct regulatory sub-networks in different cells, which helps define cell-type specific spatiotemporal gene expression patterns [3, 4].

Enhancer activity is not only dynamic throughout development, but also throughout evolutionary time [5]. The reason is that mutations in enhancer sequences can create or ablate interactions with regulatory proteins, thus enabling modifications in gene use without affecting gene product [6, 7]. Such changes alter a regulatory network's wiring diagram of "who regulates whom," which can cause changes in gene expression patterns that embody or lead to evolutionary adaptations or innovations [8]. Examples include the archetypical pentadactyl limb anatomy of extant tetrapods [9], ocular regression in subterranean rodents [10, 11], limb loss in snakes [11, 12], convergent pigmentation patterns in East African cichlids [13], the mammalian neocortex [14], and cell type diversity in eumetazoans [15].

Regulatory networks not only evolve via rewiring, but also via the addition of new genes [16]. Gene duplication, retrotransposition, gene fusion, the domestication of genomic parasites, and horizontal gene transfer are all means by which new genes can arise from pre-existing genes [17], and thus expand gene regulatory networks. In addition, it is becoming increasingly appreciated that new genes can arise *de novo* from non-coding regions of the genome [18-22]. For protein-coding genes, the essential prerequisites of this process are the formation of an open reading frame (ORF), together with the transcription and translation of that ORF. Because much of the genome is transcribed [23, 24] and many lineage-specific transcripts containing ORFs are potentially translated [25-30], the *de novo* evolution of new protein-coding genes is also a likely contributor to the growth of gene regulatory networks.

An important question concerning *de novo* genes is how they integrate into existing regulatory networks, and what role enhancers may play in this process. It has been hypothesized that enhancer acquisition allows new genes to expand their breadth of expression, providing opportunities to acquire new functions in different cellular contexts [31]. Enhancers may therefore help new genes integrate into existing regulatory networks via edge formation and rewiring. Less appreciated is the role enhancers may play in the origin of *de novo* genes [32], and thus in the growth of gene regulatory networks. The physical proximity between active enhancers and their target genes [33] – facilitated by DNA looping – creates a transcriptionally permissive environment that is engaged with RNA polymerase II, which may lead to the transcription of regions near the enhancer, or to the transcription of the enhancer itself, producing so-called enhancer RNA [1, 34]. If the resulting transcript is stable, harbors an open reading frame, and engages with ribosomes, then it fulfills the basic prerequisites of *de novo* gene birth. Thus, enhancers may play a dual role in the evolution of *de novo* genes, and consequently in the evolution of gene regulatory networks. By creating a transcriptionally permissive environment that is engaged with the transcriptional machinery, enhancers may facilitate the origin of *de novo* genes; by physically interacting with gene promoters, enhancers may facilitate the integration of *de novo* genes into existing regulatory networks.

Here, we take an integrative approach to study this potential dual role of enhancers. We leverage single-cell transcriptomic and functional genomics data from mouse that describe gene expression levels, chromatin accessibility, and chemical modifications to histones, as well as phylostratigraphic estimates of the ages of transcribed ORFs. We find that the distance between ORFs and enhancers in nucleotide sequence increases with ORF age, indicating that young ORFs preferentially emerge near enhancers. We also find that the number of enhancer interactions per ORF increases with ORF age, even across macro-evolutionary timescales. In sum, our findings support a dual role for enhancers in the origin of *de novo* genes and in their functional integration into gene regulatory networks.

Results

The maturity and age of transcribed open reading frames

To set the stage for our study, we first characterized the maturity and age of a set of mouse transcripts bearing ORFs [29]. Specifically, we characterized the transcript maturity of 46,501 murine ORFs by assessing whether *i*) the ORF resides in a region of open chromatin, which implies it is accessible to the transcriptional machinery; *ii*) the transcript has detectable 5' capping, which confers stability [35, 36], permits its export from the nucleus to the cytoplasm [37] and promotes translation [36]; and *iii*) the transcript associates with ribosomes, indicating the potential for translation [25, 29, 30]. Fig. 1A shows a schematic of our classification of transcript maturity.

We found that over a third (16,735) of the 46,501 ORFs had the highest level of transcript maturity, which we refer to as maturity level 3 (Fig. 1B). The remaining ORFs were distributed among different combinations of the three maturity indicators. We refer to ORFs found in regions of open chromatin as having a maturity level 1 (5,640 ORFs) and those that are also 5' capped as having a maturity level 2 (4,927 ORFs).

The ORFs we assessed had their phylogenetic age estimated by Schmitz et al. [29], based on their presence in the transcriptomes of other mammalian species, including rat, human, and opossum (Fig. 2A). If a homolog of a mouse ORF is found in another species, then it is assumed to have emerged before the common ancestor of that species and mouse. For example, if an ORF is shared with opossum, it is assumed to have originated before the branching of marsupials and placental mammals ~160 million years ago; if it is not shared with any of the other three species, it is assumed to have emerged only after the split between mouse and rat ~20 million years ago. Expectedly, when assessing the distribution of ORFs with each of the maturity indicators across the different age categories, we found that the older an ORF is, the more likely it is to correspond to higher levels of maturity. This is clear from the observation that the percentage of ORFs corresponding to the oldest age class (i.e., opossum) increases with the maturity level, while the percentage corresponding to the youngest age class (i.e., mouse) decreases (Fig. 2B). Furthermore, whereas most mouse-specific ORFs have a maturity level of 1, that fraction gradually decreases as ORFs grow older, while the fraction of ORFs of maturity level 3 increases with age from their minimum in mouse-specific ORFs to their maximum in opossum-shared ORFs (Fig. 2C).

Due to the resolution of the phylogeny shown in Fig. 2A, there is variation in the ages of the ORFs even within a given lineage. We therefore reasoned that such variation might be reflected by variation in transcript maturity. To determine if this was the case, we considered the expression of mouse-specific ORFs from ten different taxa from the mouse branch after the mouse-rat split (Fig. 2D) [23]. Making use of transcriptomic data from those ten taxa, we determined when in the recent phylogenetic history leading to our focal species (*Mus musculus domesticus*) did the genomic regions harboring mouse-specific ORFs start being transcribed. As anticipated, we found that whereas the fraction of non-mouse-specific ORFs with detectable transcription is relatively constant across the different lineages, fewer mouse-specific ORFs are expressed in the lineages that are more distantly related to *M. m. domesticus* (Fig. 2E). We also observed that more mature ORFs are more likely to be transcribed at more basal branches of the mouse phylogeny than are less mature ORFs, indicating that transcript maturity is indicative of when in the mouse phylogeny the genomic region harboring the ORF started being transcribed (Fig. 2F).

In sum, these results show that an ORF's transcript maturity increases with its age, complementing previous reports that focused on the correlation between age and translation potential [29]. With these estimates of transcript maturity and age at hand, we next studied the role enhancers play in the birth of *de novo* genes and in their integration into regulatory networks.

Many young and transcriptionally immature ORFs are proximal to enhancers

H3K27ac and H3K4me1 are histone modifications that are commonly used to identify enhancers, specifically when they are not found overlapping H3K4me3 modifications, which are indicative of promoters [38]. We therefore merged chromatin immunoprecipitation followed by DNA sequencing (ChIP-seq) data for H3K27ac, H3K4me1, and H3K4me3 obtained from 23 mouse tissues and cell types [39], and considered enhancers to be those genomic regions where H3K27ac and/or H3K4me1 peaks do not overlap H3K4me3 peaks in any tissue [40, 41] (Materials and Methods). Assessing the 27,347 ORFs with an assigned maturity level, we found that *i*) mouse-specific ORFs are significantly closer to enhancer marks than ORFs shared with rat, human, or opossum (Spearman's correlation coefficient ρ = 0.27, p < 0.01), with a median distance to their closest enhancer mark of 1,589bp for mouse-specific ORFs compared to more than 2,500bp for the remaining age classes (Fig. 3A); *ii*) over 30% of mouse-specific ORFs are in regions of open chromatin containing enhancer marks, while this percentage decreases as ORFs grow older to less than 5% for those shared with opossum (Fig. 3B); *iii*) significantly more enhancers are found within 50kb upstream and 50kb downstream of mouse-specific ORFs than in any other age class (Fig. S1, Wilcoxon's rank sum test p < 0.05); *iv*) the mouse-specific age class has the highest percentage of ORFs showing evidence of bidirectional transcription – a hallmark of enhancer activity [42] (Fig. 3C); and *v*) ORFs of lower transcript maturity, which tend to be younger, are nearer to enhancers than ORFs of higher transcript maturity, which tend to be older (Fig. S2). These results suggest that the birth of many new genes is facilitated by their close proximity to enhancers.

Because many (58%) of the mouse-specific ORFs are found in genomic regions that overlap or are very close to genomic regions that harbor annotated genes, we expect that at their birth, such ORFs will inherit the regulatory properties of their host gene, which is older. To specifically assess the regulatory background of ORFs that emerged from or near enhancers and thus did not coopt the regulatory features of the promoters of older genes, we separated ORFs stemming from genomic regions annotated as intergenic (which are the ORFs most likely to have emerged *de novo* [29]) from those that we considered genic, which are those ORFs overlapping other genes or that are near the promoters of other genes (Materials and Methods). We found that intergenic ORFs are considerably more likely to be found closer to enhancers than genic ORFs (Fig. 3D; Fig. S3). For example, ~65% of mouse-specific intergenic ORFs were within 1kb of an enhancer, as compared to ~25% for mouse-specific genic ORFs and ~10% for non-mouse-specific ORFs. This implies that ORFs emerging within intergenic regions of the genome lose their proximity to enhancers as they age, perhaps via the transformation of enhancers to promoters [43]. This possibility is supported by the observation that the chromatin modification indicative of promoters, H3K4me3, shows trends opposite to the ones described above for enhancers. That is, older ORFs are closer to a larger number of H3K4me3 marks than younger ORFs (Fig. S4).

These observations support the hypothesis that enhancers facilitate the *de novo* evolution of genes from non-coding DNA, and thus contribute to the expansion of gene regulatory networks. However, our analyses so far have considered enhancer marks that were merged across a diversity of cell types and tissues. To provide more direct evidence that enhancers facilitate *de novo* gene birth, we separately considered three tissues (liver, brain, and testis) for which we had both transcriptomic and histone modification data. We found that 24% (100 ORFs), 36% (931 ORFs), and 26% (244 ORFs) of intergenic mouse-specific ORFs with evidence for transcription in liver, brain, and testis, respectively, are within 1kb of an enhancer (Fig. S5). These percentages are considerably lower for genic ORFs (< 8%) and for ORFs shared with rat, human, and opossum (< 2%). Enhancers therefore provide fertile ground for the *de novo* birth of new genes from intergenic regions of the genome.

Enhancer interactions are gradually acquired over macro-evolutionary timescales

We next asked how enhancers integrate new genes into existing regulatory networks. The CCCTC-binding factor (CTCF) is an architectural DNA-binding protein that mediates physical interactions between promoters and enhancers [44]. Using ChIP-seq data for CTCF in 15 cell and tissue types, we found that CTCF-bound regions of the genome overlap a larger fraction of older ORFs than younger ORFs (~75% of opossum-shared ORFs compared to ~45% of mouse-specific ORFs; Fig. 4A), that there is a negative correlation between the age of an ORF and its distance to the closest CTCF-bound region (Spearman's correlation coefficient ρ = -0.27, $p < 0.01$), and that among young mouse-specific ORFs the distance to the closest CTCF peak is significantly higher for intergenic than genic ORFs ($p < 0.01$; Fig. S6). These results suggest that while young ORFs are proximal to enhancers, they are not specifically targeted by them. Such enhancer interactions are likely acquired gradually over time, as CTCF motifs, and other sequence changes conducive to enhancer-promoter interactions, evolve in the proximity of ORFs.

To study how ORFs acquire interactions with enhancers, we considered an enhancer-promoter interaction map derived from single-cell chromatin accessibility data in 13 murine tissues [45] (Materials and Methods). We first corroborated the negative correlation between an ORF's number of enhancer interactions and its distance to the closest CTCF-bound region (Spearman's correlation coefficient ρ = -0.35, $p < 0.01$). We then uncovered a positive correlation between the age of an ORF and its number of enhancer interactions (Spearman's correlation coefficient ρ = 0.17, $p < 0.01$; Fig. 4B). This number increased from a median of 5 enhancer interactions for mouse-specific ORFs to a median of 13 for ORFs that are shared with opossum, indicating that enhancer-promoter interactions are gradually acquired over time. However, when restricting our analysis to ORFs of the highest transcript maturity class, this positive correlation was lost (Spearman's correlation coefficient ρ = 0.001, p = 0.9).

We reasoned that this loss could be because mouse-specific ORFs of genic origin are enriched for transcripts of the highest maturity class (38% as compared to 1.4% for intergenic ORFs). We therefore partitioned the mouse-specific ORFs according to whether they were intergenic or genic, and compared the number of enhancer interactions in these classes to the number of enhancer interactions for non-mouse-specific ORFs. We found that intergenic ORFs had fewer enhancer interactions than genic ORFs, which were similar to non-mouse-specific ORFs in their number of enhancer interactions (Fig. 4C). This suggests that mouse-specific ORFs of genic origin, which are enriched for mature transcripts, tend to coopt the regulatory interactions of their host gene, or of nearby genes. To account for this confounding effect, we considered ORFs that do not share their segment of open chromatin with any other ORF and are therefore unlikely to be coopting the enhancer interactions of other genes (Materials and Methods). We call these 'single ORFs'. We use this distinction, rather than intergenic vs. genic, because only 0.06% of ORFs that emerged before the rat/mouse split are annotated as intergenic, whereas 48% can be considered single ORFs. After making this distinction, we recovered the positive correlation between an ORF's number of enhancer interactions and its age (Spearman's correlation coefficient $\rho = 0.24$, $p < 0.01$); even for ORFs of the highest transcript maturity class, we found that mouse-specific ORFs were involved in fewer interactions than opossum-shared ORFs (Wilcoxon's tailed test, $p < 0.01$; Fig. 4D). Therefore, intergenic mouse-specific ORFs with the highest level of transcript maturity, which tend to be older than those with lower levels of transcript maturity (Fig. 2), have fewer interactions than ORFs in the oldest age class, providing further evidence of the gradual acquisition of enhancer interactions over time.

To further explore the pace at which new enhancer interactions are gained over evolutionary time, we shifted our focus to opossum-shared ORFs, most of which (95%) correspond to annotated genes. We separated these into 15 new age classes dating back to the origin of cellular life [46] in order to understand how enhancer interactions are acquired over macroevolutionary timescales (Fig. 5A). With the sole exception of the oldest genes shared with bacteria and archaea, which have significantly fewer interactions than ORFs that emerged before the common ancestor of all eukaryotes, no other age class shows significantly fewer interactions than a younger age class (Fig. 5B; in Fig. S7, note that only a single element below the main diagonal is significant). Disregarding ORFs from the oldest age class, we found a significant correlation between the age of genes and their number of enhancer interactions (Spearman's correlation coefficient $\rho = 0.15$, $p < 0.01$).

In sum, young ORFs have relatively few interactions with enhancers, despite being proximal to them in nucleotide sequence. As ORFs age, they gradually acquire enhancer interactions (Fig. 4), a process that continues over macroevolutionary timescales (Fig. 5B).

Enhancer acquisition influences expression breadth and variance

We next explored the functional consequences of enhancer acquisition. To do so, we first studied the expression breadth of opossum-shared annotated genes using the phylogeny shown in Fig. 5A and single-cell transcriptomic data from 68 cell types of ten murine tissues [47], for which we also had single-cell chromatin accessibility data (Materials and Methods). We found that expression breadth increases with gene age (Spearman's correlation coefficient $\rho = 0.30$, $p < 0.01$; Fig S8A), corroborating previous analyses performed using transcriptomic data from whole tissues [48]. We additionally found that a gene's expression breadth increases with its number of enhancer interactions (Spearman's correlation coefficient $\rho = 0.37$, $p < 0.01$; Fig. 5C), suggesting that enhancer acquisition has functional consequences.

We next measured the coefficient of variation for the expression of each gene, a measure that is useful for identifying stably vs. variably expressed genes from single cell RNA sequencing [49]. It is calculated as the standard deviation of a gene's expression across cell types, divided by the mean expression across cell types (Materials and Methods). Genes with a lower coefficient of variation tend to be more tightly regulated than those with a higher coefficient of variation [49]. We found a significant correlation between the coefficient of variation and gene age (Spearman's correlation coefficient ρ = -0.32, p < 0.01; Fig. S8B), as well as with a gene's number of enhancer interactions (Spearman's correlation coefficient ρ = -0.32, p < 0.01; Fig 5D). Specifically, the coefficient of variation decreases as genes acquire more enhancer interactions, stabilizing around one when genes acquire at least 20 enhancer interactions. These results show that enhancer acquisition affects gene expression breadth and variance, further supporting the role of enhancers in the integration of genes into regulatory networks.

Discussion

We report a dual role of enhancers in the evolution of gene regulatory networks: They engage with the transcriptional machinery to create an environment of open chromatin that is conducive to the *de novo* birth of new genes, and they help integrate these new genes into existing regulatory networks by interacting with gene promoters, thus facilitating the evolution of controlled and robust gene expression in space and time.

Our study provides empirical support for the hypothesis that enhancers may facilitate *de novo* gene evolution, which to our knowledge was first proposed upon the discovery of enhancer RNA [34] and later expanded upon in a perspective piece by Wu and Sharp [32]. Our findings complement contemporaneous work [50] on the regulatory architecture of the nematode *Pristionchus pacificus*, which showed that young genes – those private to *P. pacificus* – are in closer proximity to enhancers than genes with one-to-one orthologs in other nematode species. The observation that enhancers facilitate *de novo* gene birth in both nematodes and mammals suggests that this mode of *de novo* gene evolution dates back to at least the common ancestor of Bilateria, and possibly even earlier, since cnidarians and ctenophores also employ distal regulatory elements [15, 51, 52].

The facilitating role of enhancers in *de novo* gene birth is conceptually similar to the facilitating role of the permissive chromatin state of meiotic spermatocytes and post-meiotic round spermatids that underlies the "out-of-testis hypothesis," which proposes the testis as a primary tissue for the origination of new genes [17]. Both scenarios envision regions of open chromatin that are exposed to the transcriptional machinery, and thus produce a transcriptionally active environment that is conducive to the evolution of new genes. The two scenarios differ, however, in at least two ways. First, genes that emerge from or near enhancers may rapidly acquire their own promoters, due to the similar architectural and functional features of enhancers and promoters, a similarity that facilitates the rapid turnover of the former to the latter [43]. Second, enhancers are often deployed in multiple cell types or developmental stages [53], exposing enhancer-proximal *de novo* genes to distinct cellular contexts where they may confer a selective advantage.

The hypothesis that enhancers help *de novo* genes integrate into existing regulatory networks was previously proposed in the context of the out-of-testis hypothesis, as a means to expand a new gene's breadth of expression [31]. Using single-cell chromatin accessibility and transcriptomic data, our study provides the first empirical support for the hypothesis that *de novo* genes gradually acquire enhancer interactions over time, and that this acquisition increases expression breadth. These findings complement related studies of gene integration into cellular networks, such as networks of protein-protein interactions [54, 55]. Our observation that genes continue to acquire enhancer interactions over macro-evolutionary timescales mirrors similar increases in other aspects of gene regulation, such as in the number of proximal transcription factor binding sites, alternative transcript isoforms, and miRNA targets [56].

Regulatory networks drive the spatiotemporal gene expression patterns that give rise to and define the numerous and distinct cellular identities characteristic of Metazoan life. Enhancers play an integral role in this process, mediating cell-type-specific gene-gene interactions, thus facilitating the combinatorial deployment of different genes in different contexts. Genetic changes that affect such interactions are responsible for myriad evolutionary adaptations and innovations [6-8, 57]. Our results suggest that the power of enhancers in creating such evolutionary novelties lies not only in their ability to rewire gene regulatory networks, but also in their ability to expand them, by providing fertile ground for *de novo* gene birth.

Materials and methods

ORF age and transcript maturity

Schmitz et al. [29] identified a set of 58,864 ORFs from the transcriptomes of three murine tissues: liver, testis, and brain. Blasting against the transcriptomes of four other mammalian species (rat, human, kangaroo rat, and opossum), they estimated the age of each ORF by phylostratigraphic methods [29, 58]. Because of the small number of ORFs shared with the kangaroo rat (49 ORFs), we merged these ORFs together with those from the rat age class. We used the genomic coordinates of the first exon of each ORF in the mm10 mouse genome reference to study the regulatory properties of ORFs of different ages, for example to study their distance to the nearest enhancer.

We considered three indicators of ORF transcript maturity:

i) Open chromatin: We used single-cell ATAC-seq data from 13 different mouse tissues (bone marrow, cerebellum, large intestine, heart, small intestine, kidney, liver, lung, cortex, spleen, testes, thymus, and whole brain). The ATAC-seq method detects regions of open chromatin through the insertion of transposons in random accessible regions of the genome that can later be sequenced [59]. We obtained the data from the Mouse ATAC atlas [45], which comprised 436,206 peaks of open chromatin. We used liftOver from the Genome Browser at UCSC [60] to convert the genome coordinates from mm9 to mm10. A total of 29 peaks could not be converted. Using the "intersect" function of bedtools with default parameters [61], we found which ORFs have their first exons in regions of open chromatin and are therefore accessible to the transcriptional machinery in at least one of the tissues.

ii) 5' capping: We used cap analysis of gene expression (CAGE) data from the FANTOM5 consortium from 1,016 mouse samples including cell lines, primary cells and tissues [62, 63]. This method is based on the capture of 5' capped ends of mRNA, which allows the mapping of regions of transcription initiation genome-wide [64]. Using the "closest" function from bedtools with default parameters [61], we measured the distance between an ORF's first exon and its closest CAGE peak. We considered a transcript to be 5' capped if the start site of its first exon was located within 200 bases of a CAGE peak (Fig. S9).

iii) Ribosome association: We used ribosome profiling (ribo-seq) data from 9 different mouse tissues (embryonic stem cells, neutrophils, fibroblasts, liver, brain, testis, epidermis, kidney, and adipose tissue). This method is based on the sequencing of mRNA fragments that are protected from RNase digestion by ribosomes [65]. We obtained the coordinates of mRNA segments detected by ribo-seq from GWIPS-viz [66], a database that includes such data from different studies. Following Schmitz et al. [29], we considered an ORF as being potentially translated if at least one read from the ribo-seq datasets could be assigned to the ORF in question.

Using these indicators, we defined three levels of transcript maturity: maturity level 1 for ORFs whose first exon overlaps open chromatin, maturity level 2 for ORFs that are also 5' capped, and maturity level 3 for ORFs that also associate to ribosomes. Because the ribo-seq data may be limited by the detectability of the transcript [29], we only considered ORFs that were also found in the mRNA-seq dataset available at GWIPS-viz; this filter lead us to only consider a subset of the ORFs reported by Schmitz et al. [29]. Specifically, we assigned transcript maturity levels to 46,501 ORFs (~79% of the 58,864 ORFs).

To determine if transcript maturity correlates with gene age even within the mouse lineage, we considered the transcriptomes of brain, liver and testis from 10 different mouse taxa (3 populations of *Mus musculus domesticus*, 2 populations of *M. m. musculus*, and 1 from *M. m. castaneus*, *M. spicilegus*, *M. spretus*, *M. mattheyi* and *Apodemus uralensis*). The data consisted of read counts from the transcriptomes of each taxon mapped to 200 bp windows of the mm10 mouse reference genome [23]. We considered an ORF to be expressed in any of the ten taxa if at least 10 reads (the upper threshold to be considered "lowly expressed" [23]) could be detected in the 200 bp windows overlapping at least 60% of the length of the first exon of the ORF.

Enhancer association

We obtained ChIP-seq data for H3K27ac, H3K4me1, and H3K4me3 modifications from 23 different tissues and cell types from the ENCODE project (bone marrow, cerebellum, cortex, heart, kidney, liver, lung, olfactory bulb, placenta, spleen, small intestine, testis, thymus, embryonic whole brain, embryonic liver, embryonic limb, brown adipose tissue, macrophages, MEL, MEF, mESC, CH12 cell line, and E14 embryonic mouse) [39]. We used liftover to convert the genomic coordinates of the peaks from mm9 to mm10. We used the "merge" function of bedtools with default parameters to collate the peaks for all tissues and cell types, considering any overlapping H3K27ac and H3K4me1 peak as part of the same enhancer. We used the "intersect" function of bedtools with default parameters to separate H3K27ac and H3K4me1 peaks that overlapped any length of H3K4me3 peaks from those that did not. This resulted in 172,930 H3K27ac and 277,187 H3K4me1 peaks that did not overlap H3K4me3 peaks. We considered genomic regions with H3K4me3 peaks to be promoters, and those exclusively with H3K27ac and/or H3K4me1 peaks to be enhancers [41]. We measured the distance in base pairs between the first exon of an ORF to an enhancer or promoter using the "closest" function of bedtools with default parameters. To assess the number of enhancers surrounding an ORF, we considered the 50,000 base pairs upstream and downstream of the first exon of each ORF, and determined the number of H3K27ac and H3K4me1 peaks within that window.

We also studied the association of ORFs that are expressed in different tissues to chromatin modifications in those same tissues. To do so, we used the transcriptomic data for brain, testis and liver from the samples of *Mus musculus domesticus* as described in the previous section to classify ORFs as expressed or not expressed in each tissue. We determined the fraction of ORFs expressed in each tissue that were up to 1kb away from a H3K4me1, H3K27ac and H3K4me3 ChIP-seq peak identified from liver, testis, embryonic whole brain, and cortex samples.

We also considered bidirectional CAGE peaks, which are indicative of enhancers [42, 67]. We assigned bidirectional CAGE peaks to ORFs using the same criteria we used to assign H3K27ac and H3K4me1 peaks to ORFs, as described above.

ORF origin

Schmitz et al. [29] annotated each ORF as belonging to one of 8 different categories: "intergenic," "close to promoter same strand," "close to promoter opposite strand," "overlapping same strand," "overlapping opposite strand," "overlapping coding sequence same strand," "overlapping coding sequence opposite strand," and "overlapping annotated gene in frame." We considered all categories except "intergenic" to be "genic" in order to separate ORFs that are born within or near existing genes from those that are not. This classification is more challenging for non-mouse-specific ORFs due to the better annotation of older genes [29], which makes them more likely to correspond to the "overlapping annotated gene in frame" category even if they are of intergenic origin. We therefore further classified ORFs according to whether they shared their segment of open chromatin with another ORF. Specifically, we classified an ORF as "shared" if its first exon was in the same segment of open chromatin as the first exon of any other ORF, and as "single" otherwise.

Enhancer interactions

As with H3K27ac, H3K4me1, and H3K4me3 histone modifications, we evaluated the distance of each ORF to CTCF ChIP-seq peaks obtained from 15 different cell and tissue types (bone marrow, cerebellum, cortex, heart, kidney, developing limb during stage E14.5, liver, fibroblasts, mESC, olfactory bulb, small intestine, spleen, testis, thymus and the whole brain) [39]. We used liftOver to convert the data from mm9 to mm10.

Cusanovich et al. [45] used single-cell ATAC-seq data to predict physical interactions between regions of open chromatin [68], thus creating an atlas of enhancer interactions in single murine cells. We downloaded these data from the Mouse ATAC atlas [45], which includes the cell clusters where the interactions occur, as well as the co-accessibility scores of pairs of regions of open chromatin – a measure of interaction strength. We disregarded cell clusters classified as "unknown" or "collisions", as well as interactions with a co-accessibility score lower than 0.25, following Pliner et al. [68]. We also filtered out interactions with regions of open chromatin that harbored annotated promoters, in order to focus solely on interactions with enhancers. An interaction was assigned to an ORF if the ORF's first exon was included in the interaction.

Age of annotated genes

To study how genes acquire enhancer interactions over macro-evolutionary timescales, we considered the subset of ORFs that belong to the opossum age class in Schmitz et al. [29] and that are annotated as genes in the latest version of Ensembl (release 95) [69]. We matched these genes to age estimates reported by Neme & Tautz [46], based on a phylostratigraphic analysis of 20 lineages spanning 4 billion years from the last universal common ancestor to the common ancestor of mouse and rat. We further filtered the dataset to only include ORFs that emerged in the first 15 of the 20 phylostrata, in order to focus on ORFs that are considered to have emerged before the split between the common ancestor of placental mammals and marsupials by both Schmitz et al. [29] and Neme & Tautz [46]. This left us with ~16,000 ORFs corresponding to annotated genes that emerged prior to the origin of placental mammals.

Breadth of expression

To study the transcription of annotated genes, we used the expression data reported by the Tabula Muris Consortium [47] for the single-cell RNA sequencing performed with FACS-based cell capture in plates, for 20 different mouse tissues. The data include the log-normalization of 1 + counts per million for each of the annotated genes in each of the sequenced cells. We considered ten tissues that were also used for the construction of the Mouse ATAC Atlas [45]. We measured the expression breadth of each ORF corresponding to an annotated gene as the number of cell types in which expression could be detected in at least 5% of the cells assigned to a cell type. Additionally, we calculated the coefficient of variation of the expression of each gene as the standard deviation over the mean of the log-normalisation of 1 + counts per million across cell types.

Figures

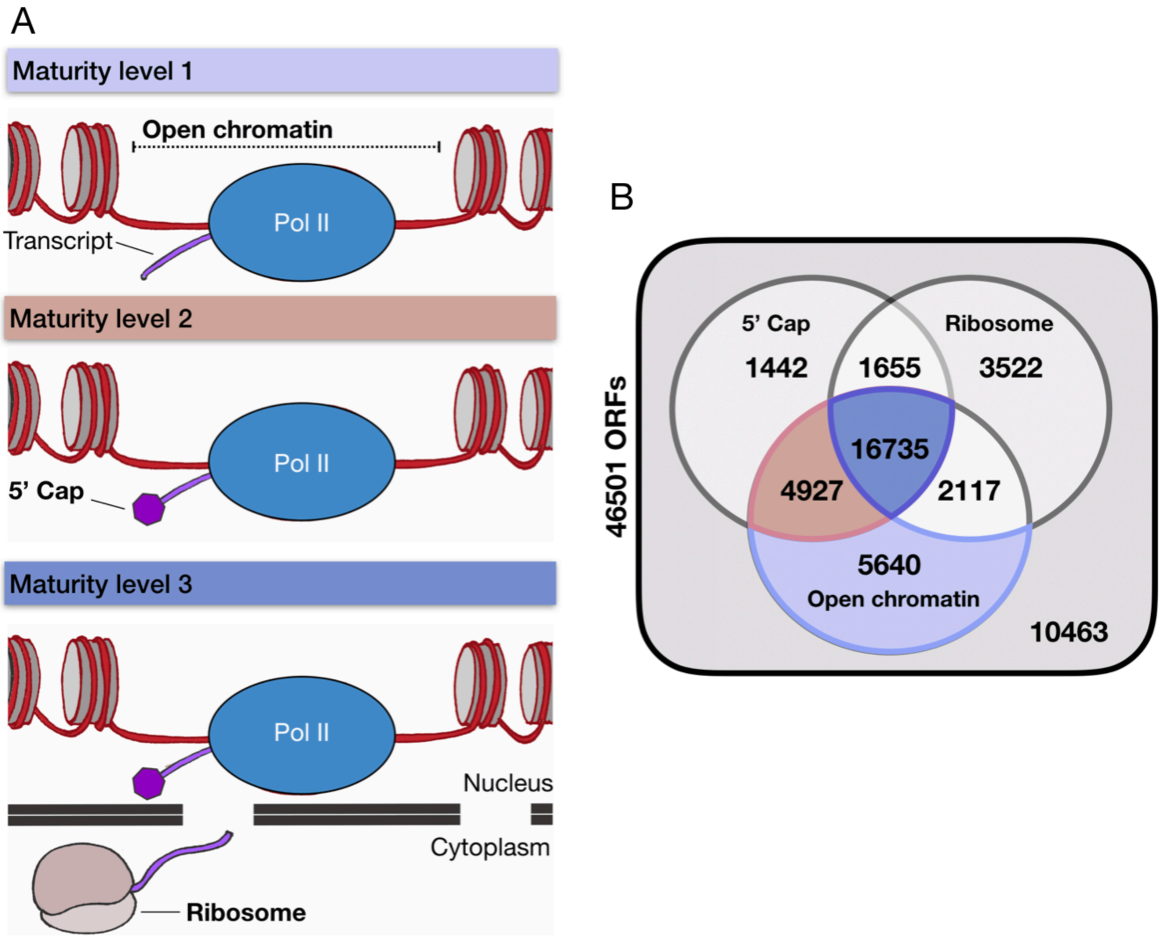

Figure 1. Three levels of transcript maturity. A) Maturity level 1 refers to ORFs that are in regions of open chromatin, but have none of the other maturity indicators; ORFs of maturity level 2 are in regions of open chromatin and are 5' capped, but have no evidence of association with ribosomes; ORFs of maturity level 3 are in regions of open chromatin, are 5' capped, and show evidence of association with ribosomes. B) Venn diagram of the number of ORFs associated with each maturity indicator. Colors correspond to the pallet used in A.

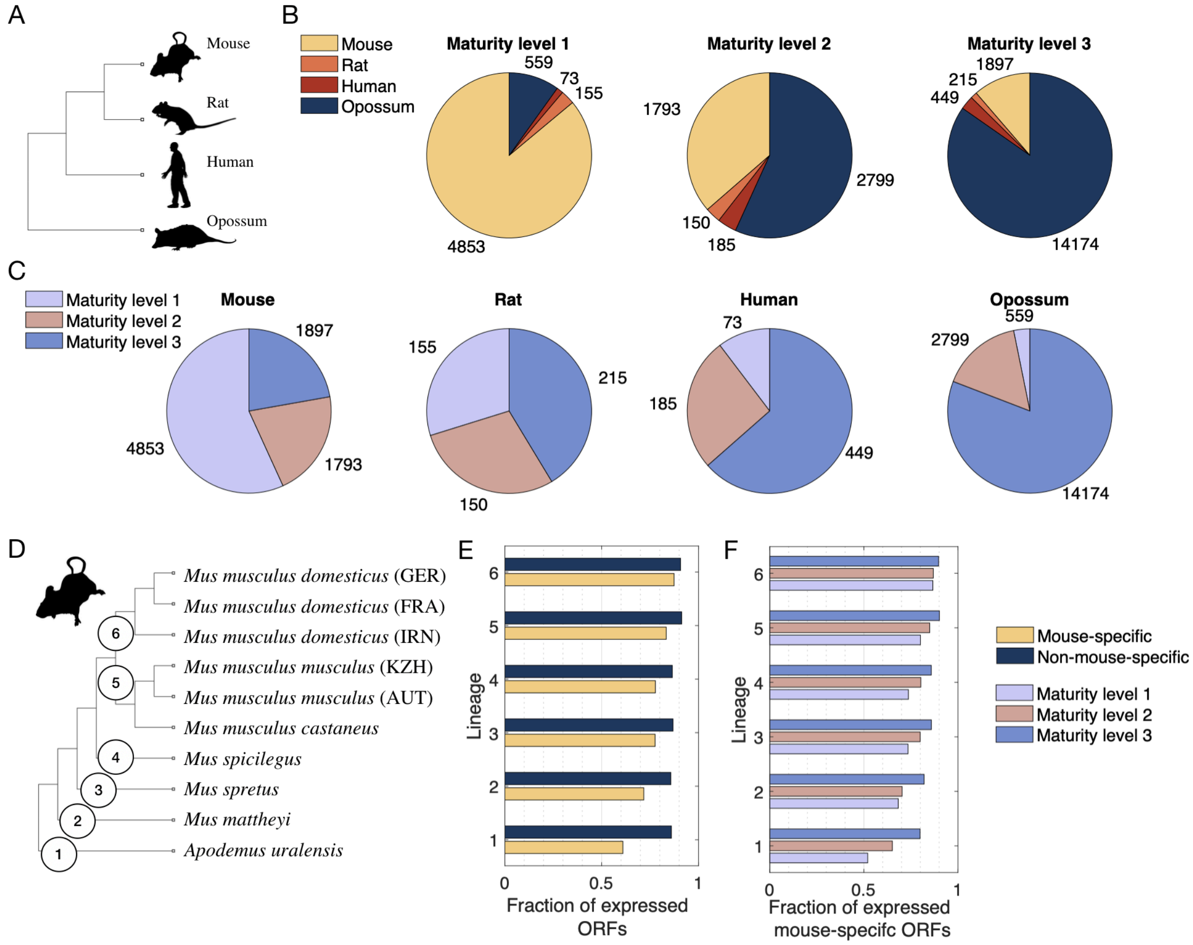

Figure 2. Transcript maturity level and ORF age. A) Phylogenetic relationship between mouse, rat, human, and opossum – the four species defining each age class. B) Pie charts of the distribution of ORFs from each maturity level among the different age classes. C) Pie charts of the distribution of ORFs from each age class among the different maturity levels. D) Phylogeny adapted from Neme & Tautz [23] of ten mouse taxa used to study the association between the transcription and the maturity level of mouse-specific ORFs. The numbered circles indicate the mouse lineages used for transcriptomic comparisons. E) Fraction of mouse-specific and non-mouse-specific ORFs for which there is evidence of transcription in brain, testis and/or liver in at least one of the taxa included in each of the six mouse lineages. F) Fraction of mouse-specific ORFs of each maturity level with detectable transcription in at least one of the taxa included in each of the six mouse lineages.

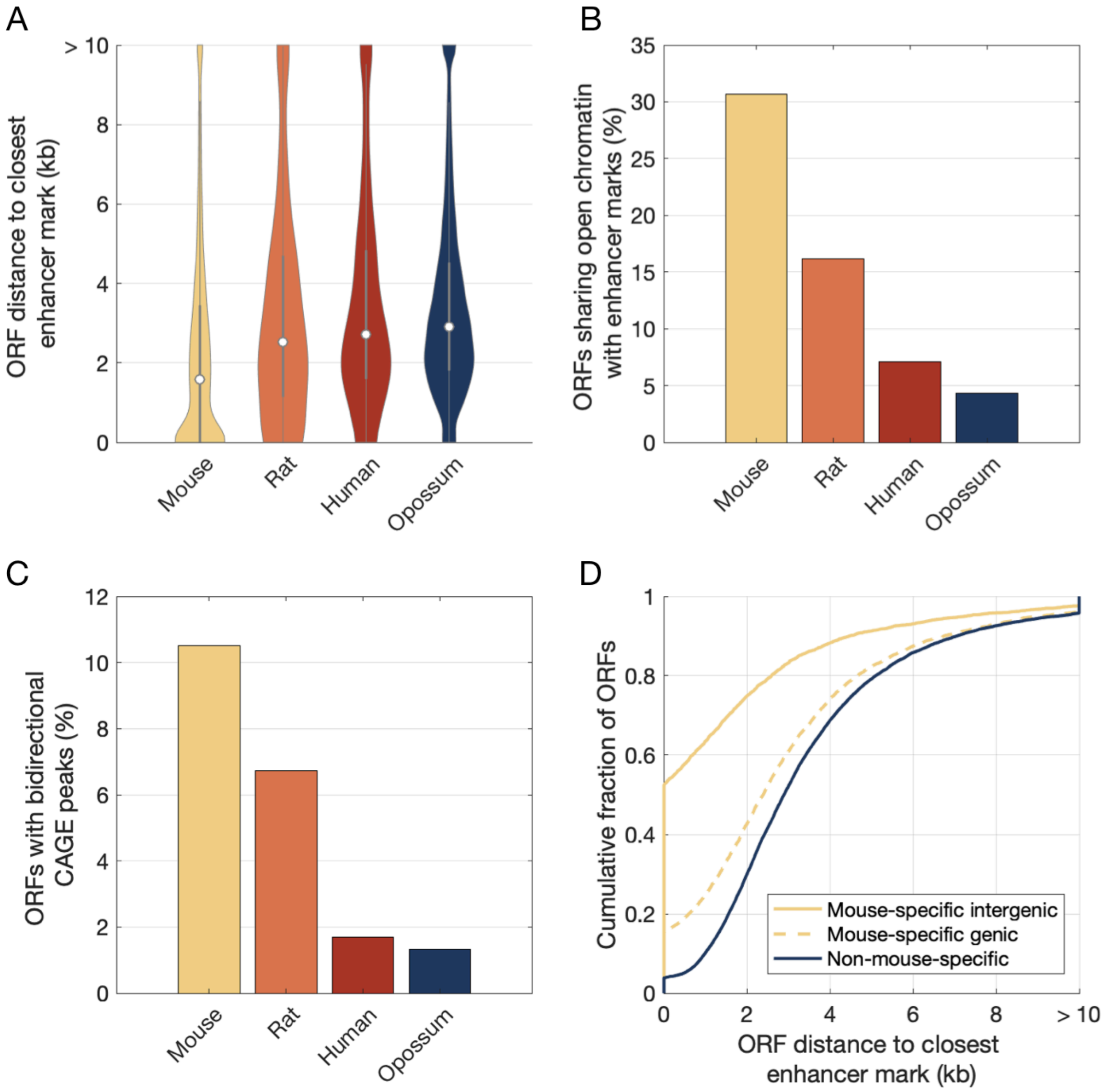

Figure 3. Enhancers facilitate *de novo* gene birth. A) Distance between each ORF and its closest H3K27ac and/or H3K4me1 peak, as a function of ORF age. B) Fraction of ORFs of each age class that share their segment of open chromatin with an enhancer mark. C) Fraction of ORFs from each age class that are within 200bp of a CAGE peak that is annotated as bidirectional [67]. D) Cumulative fraction of mouse-specific ORFs from genic and intergenic regions, as well as non-mouse-specific ORFs, as a function of their proximity to enhancer marks.

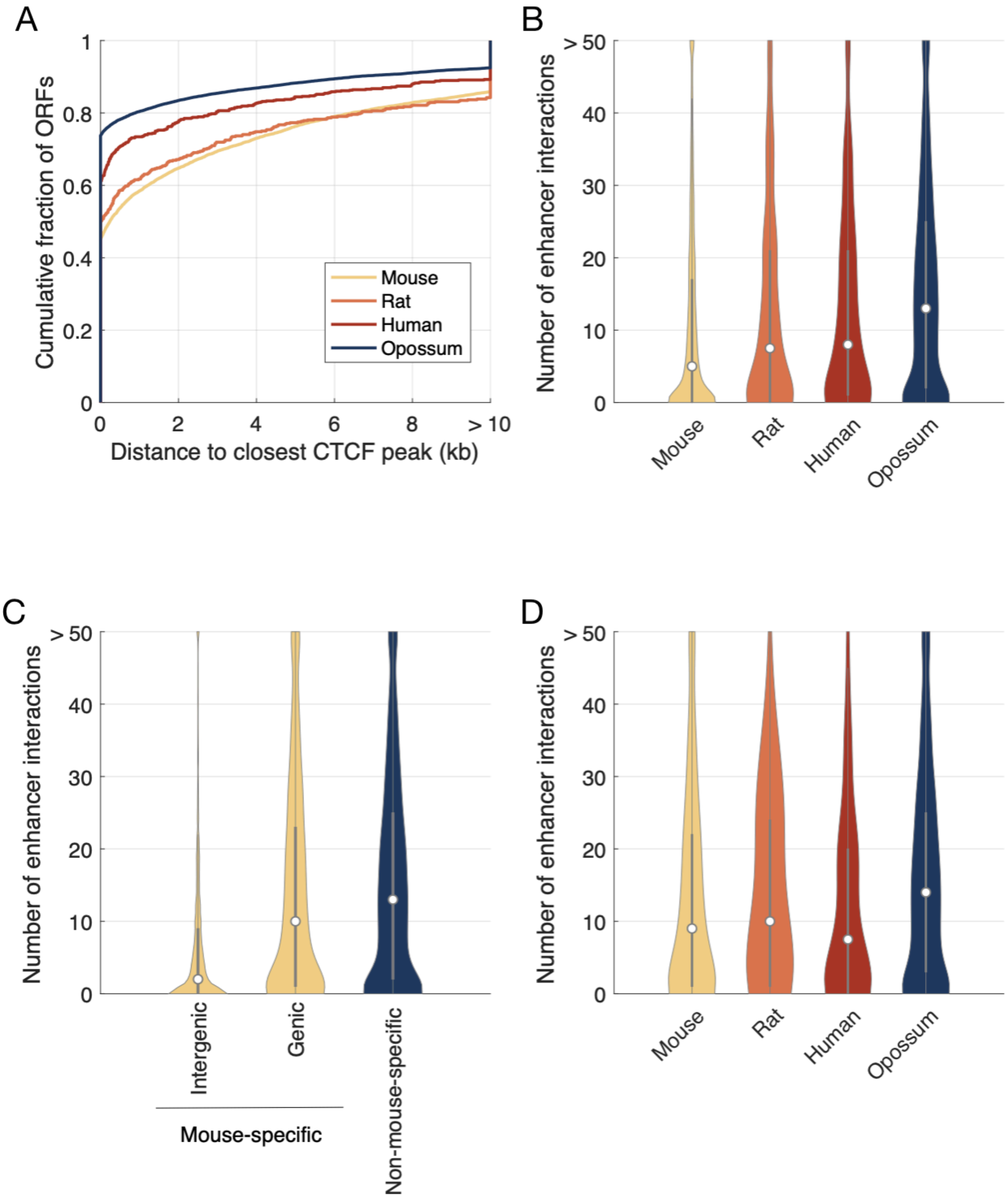

Figure 4. The number of enhancer interactions increases with ORF age. A) Cumulative fraction of ORFs of each age class as a function of their distance to the closest CTCF peak. B) Number of enhancer interactions of ORFs from each age class. C) Number of enhancer interactions of non-mouse-specific, mouse-specific genic, and mouse-specific intergenic ORFs. D) Number of interactions of single ORFs of maturity level 3 from each age class.

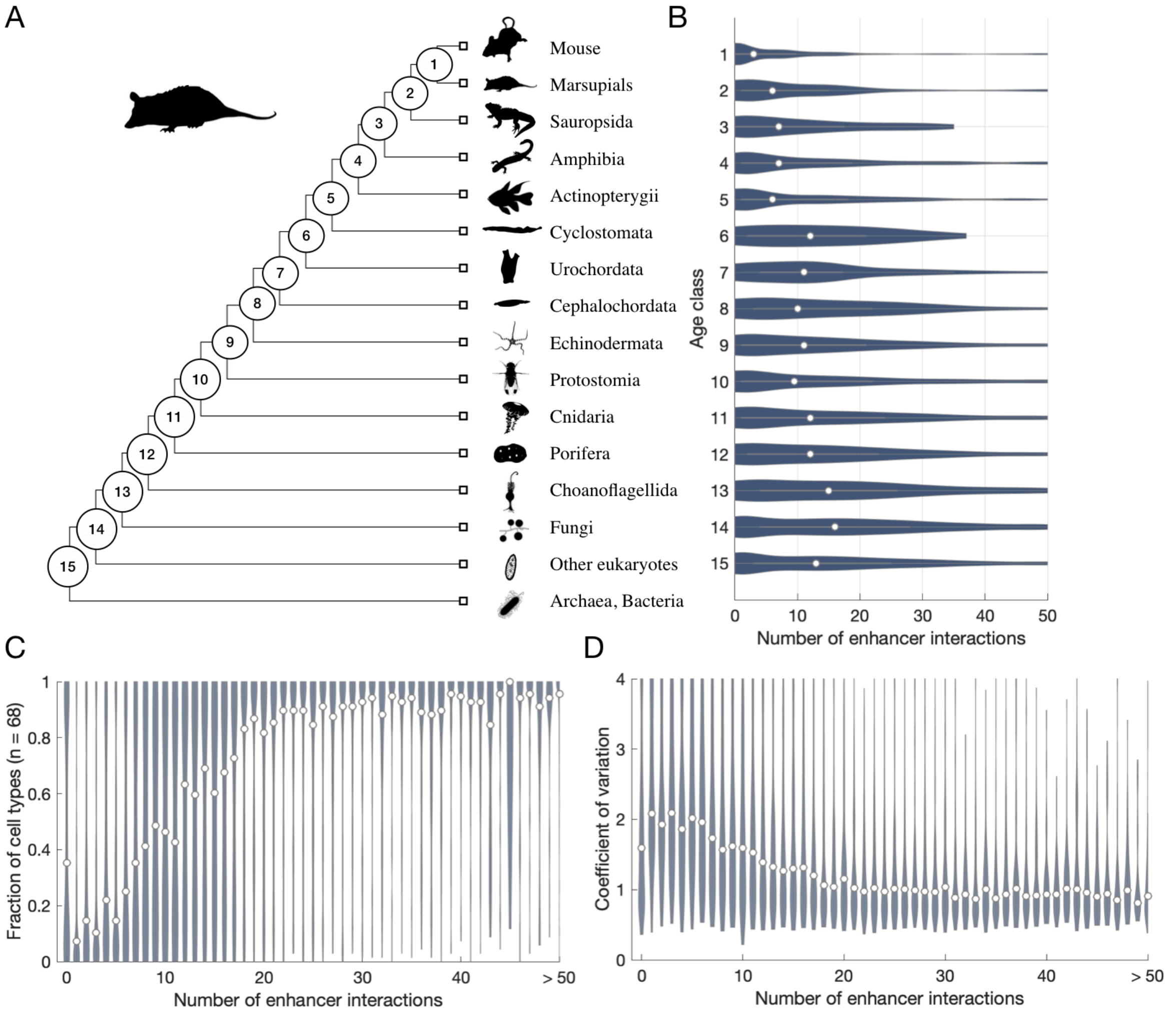

Figure 5. Enhancers facilitate the functional integration of genes into regulatory networks across macroevolutionary timescales. A) Phylogeny adapted from [46]. The numbered circles indicate lineages representative of the age classes to which genes were assigned. B) Number of enhancer interactions per gene as a function of gene age. C) Expression breadth and D) coefficient of variation as a function of the number of enhancer interactions.

Supplementary figures

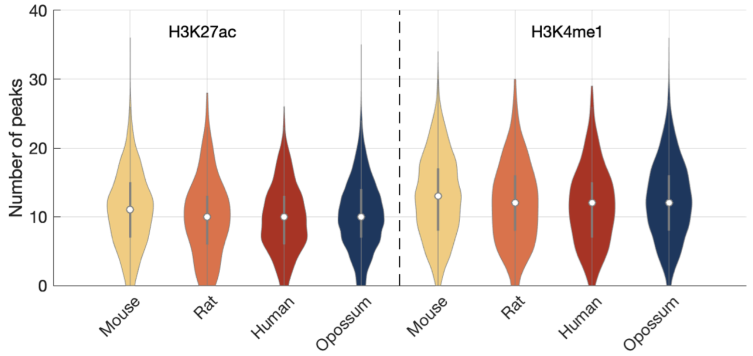

Figure S1. More enhancers are found near mouse-specific ORFs than are found near older ORFs. The number of H3K27ac and H3K4me1 peaks flanking ORFs within 50kb upstream and 50kb downstream is shown as a function of ORF age.

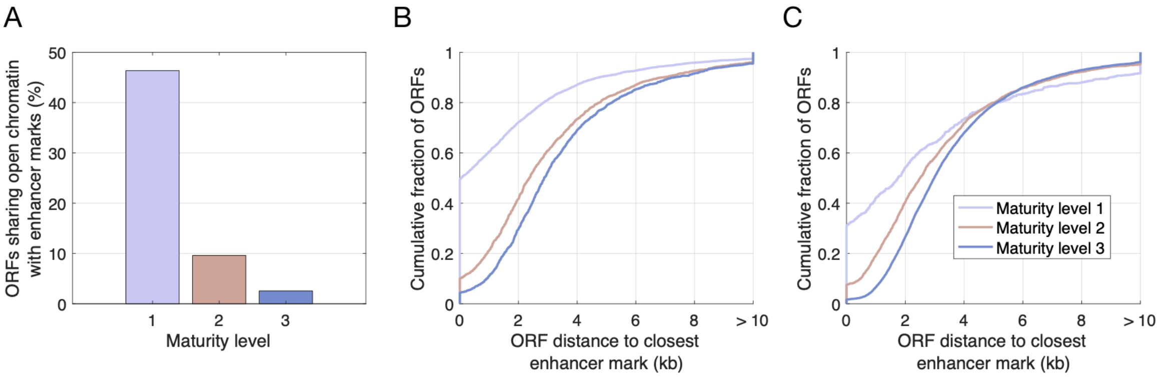

Figure S2. Distance to enhancers increases with transcript maturity. A) Fraction of ORFs of each maturity level that share their segment of open chromatin with an H3K27ac and/or H3K4me1 peak. Cumulative fraction of B) mouse-specific and C) non-mouse specific ORFs classified according to their maturity level, as a function of their proximity to the closest enhancer mark.

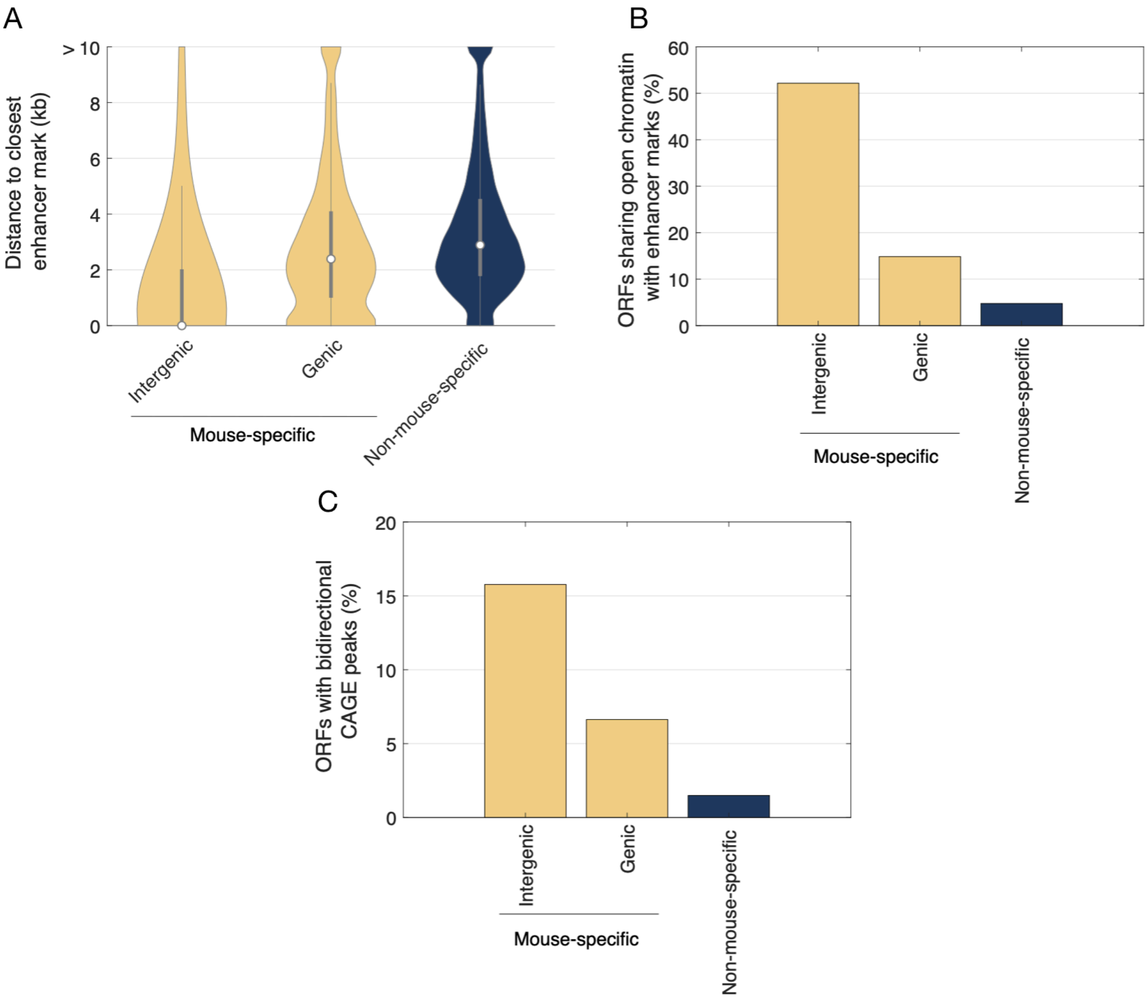

Figure S3. Mouse-specific ORFs transcribed from intergenic regions are close to enhancers. A) Distance between each ORF and its closest H3K27ac and/or H3K4me1 peak, as a function of the genomic annotation of each ORF. B) Fraction of intergenic, genic and non-mouse-specific ORFs that share their segment of open chromatin with an enhancer mark. C) Fraction of intergenic, genic and non-mouse-specific ORFs that are within 200bp of a CAGE peak that is annotated as bidirectional [67].

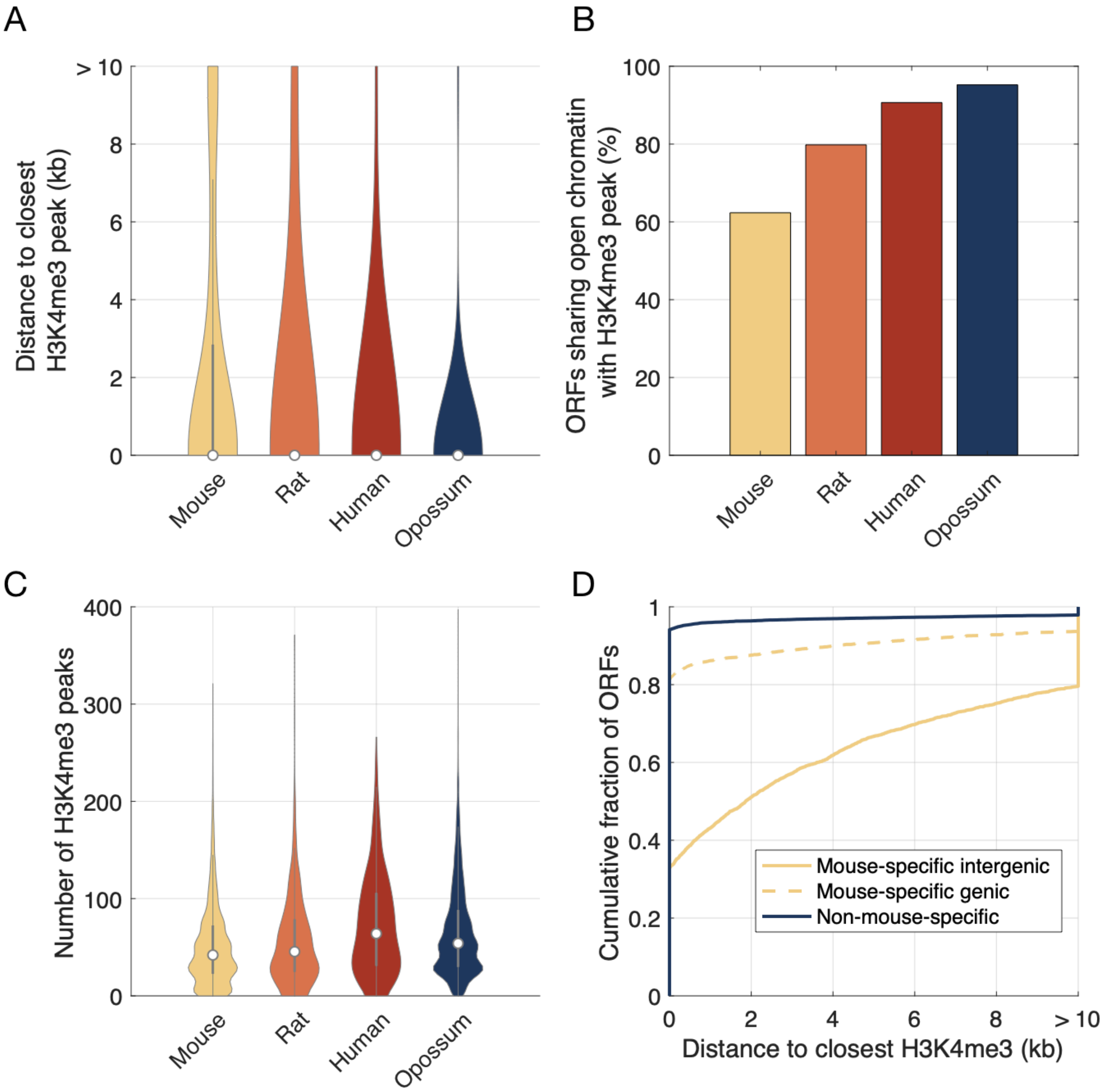

Figure S4. Older ORFs are nearer to promoters than younger ORFs. A) Distance between each ORF and its closest H3K4me3 peak, as a function of ORF age. B) Fraction of ORFs of each age class that share their segment of open chromatin with an H3K4me3 mark. C) Number of H3K4me3 peaks within 50 kb upstream or downstream of an ORF, as a function of ORF age. D) Cumulative fraction of mouse-specific ORFs from genic (dashed yellow line) and intergenic (solid yellow line) genomic regions, as well as non-mouse-specific ORFs (blue line), as a function of their proximity to H3K4me3 peaks.

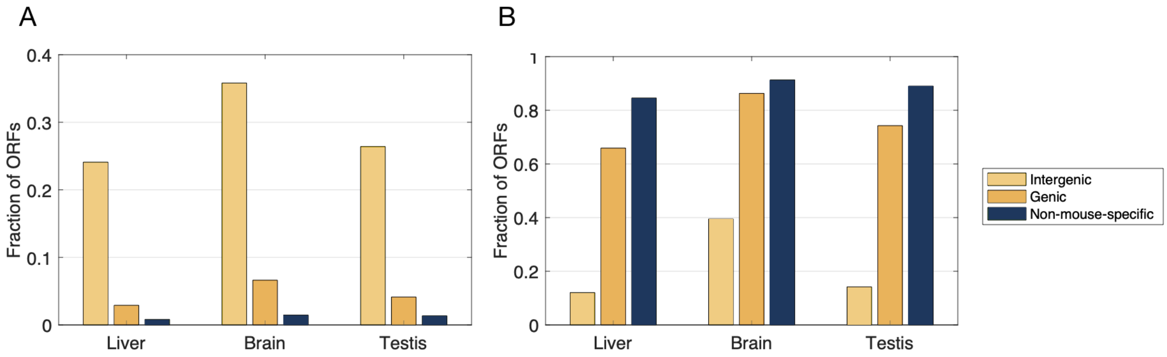

Figure S5. Intergenic ORFs preferentially emerge near enhancers. Fraction of ORFs expressed in liver, brain, and testis that are within 1kb of an active A) enhancer mark (i.e., H3K27ac or H3K4me1) or B) promoter mark (i.e., H3K4me3) in each tissue.

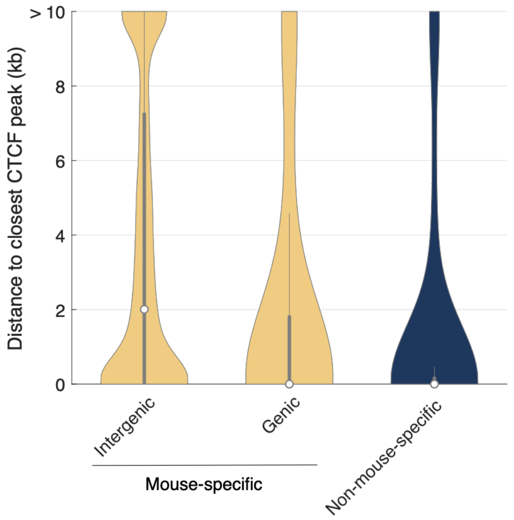

Figure S6. Intergenic ORFs are farther away from CTCF-bound regions. Distance between each ORF and its closest CTCF peak for intergenic, genic and non-mouse-specific ORFs.

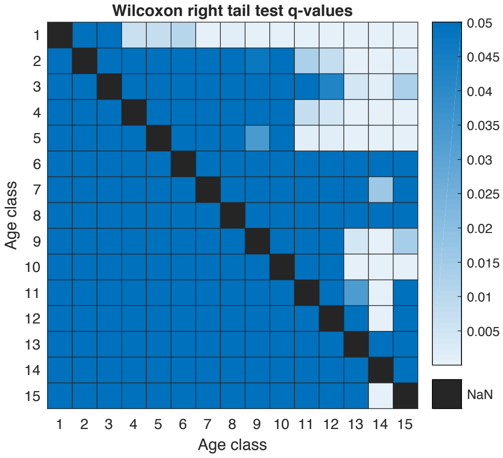

Figure S7. Heatmap of FDR-corrected q-values for the Wilcoxon right-tailed test between the number of distal interactions from each pair of age classes. Darker colors indicate higher q-values. All comparisons with a value greater than or equal to 0.05 are the darkest shade and are considered non-significant.

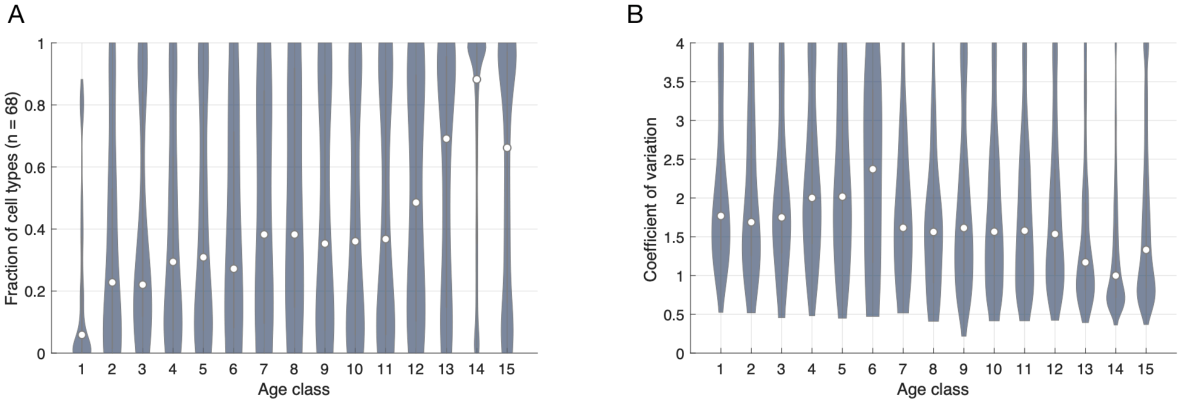

Figure S8. Expression breadth and variance correlate with gene age. A) Fraction of cell types in which there is detectable expression of annotated genes (in at least 5% of the cells included in the cell type cluster) as a function of gene age. B) Coefficient of variation of annotated genes as a function of gene age.

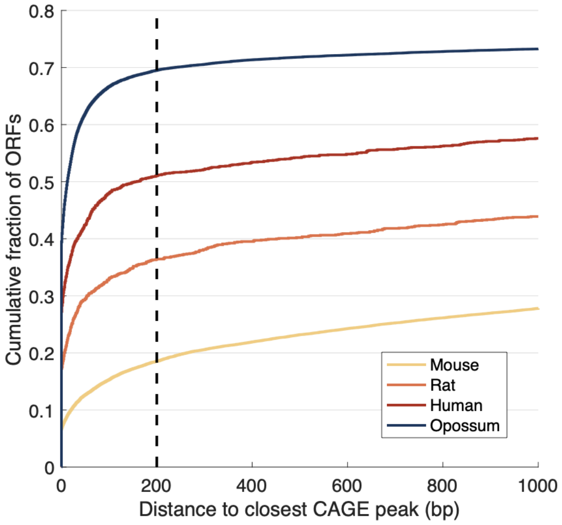

Figure S9. Cumulative fraction of ORFs of each age class as a function of the distance to the closest CAGE peak. The vertical dashed line indicates the threshold we used to consider an ORF as 5' capped because of its proximity to a CAGE peak.

References

1. Haberle V, Stark A. Eukaryotic core promoters and the functional basis of transcription initiation. Nat Rev Mol Cell Biol. 2018;19(10):621-37. Epub 2018/06/28. doi: 10.1038/s41580-018-0028-8. PubMed PMID: 29946135; PubMed Central PMCID: PMCPMC6205604.
2. Catarino RR, Stark A. Assessing sufficiency and necessity of enhancer activities for gene expression and the mechanisms of transcription activation. Genes & Development. 2018;32(3-4):202-23. doi: 10.1101/gad.310367.117.
3. Davidson EH, Levine MS. Properties of developmental gene regulatory networks. Proc Natl Acad Sci U S A. 2008;105(51):20063-6. Epub 2008/12/24. doi: 10.1073/pnas.0806007105. PubMed PMID: 19104053; PubMed Central PMCID: PMCPMC2629280.
4. Spitz F, Furlong EEM. Transcription factors: from enhancer binding to developmental control. Nat Rev Genet. 2012;13(9):613-26. doi: 10.1038/nrg3207.
5. Villar D, Berthelot C, Aldridge S, Rayner Tim F, Lukk M, Pignatelli M, et al. Enhancer Evolution across 20 Mammalian Species. Cell. 2015;160(3):554-66. doi: 10.1016/j.cell.2015.01.006.
6. Prud'homme B, Gompel N, Carroll SB. Emerging principles of regulatory evolution. Proceedings of the National Academy of Sciences. 2007;104(Supplement 1):8605-12. doi: 10.1073/pnas.0700488104.
7. Carroll SB. Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. Cell. 2008;134(1):25-36. doi: 10.1016/j.cell.2008.06.030.
8. Peter Isabelle S, Davidson Eric H. Evolution of Gene Regulatory Networks Controlling Body Plan Development. Cell. 2011;144(6):970-85. doi: 10.1016/j.cell.2011.02.017.
9. Kherdjemil Y, Lalonde RL, Sheth R, Dumouchel A, de Martino G, Pineault KM, et al. Evolution of Hoxa11 regulation in vertebrates is linked to the pentadactyl state. Nature. 2016;539(7627):89-92. doi: 10.1038/nature19813.
10. Partha R, Chauhan BK, Ferreira Z, Robinson JD, Lathrop K, Nischal KK, et al. Subterranean mammals show convergent regression in ocular genes and enhancers, along with adaptation to tunneling. eLife. 2017;6. doi: 10.7554/eLife.25884.
11. Roscito JG, Sameith K, Parra G, Langer BE, Petzold A, Moebius C, et al. Phenotype loss is associated with widespread divergence of the gene regulatory landscape in evolution. Nature Communications. 2018;9(1). doi: 10.1038/s41467-018-07122-z.
12. Kvon EZ, Kamneva OK, Melo US, Barozzi I, Osterwalder M, Mannion BJ, et al. Progressive Loss of Function in a Limb Enhancer during Snake Evolution. Cell. 2016;167(3):633-42.e11. doi: 10.1016/j.cell.2016.09.028.
13. Kratochwil CF, Liang Y, Gerwin J, Woltering JM, Urban S, Henning F, et al. Agouti-related peptide 2 facilitates convergent evolution of stripe patterns across cichlid fish radiations. Science. 2018;362(6413):457-60. doi: 10.1126/science.aao6809.
14. Emera D, Yin J, Reilly SK, Gockley J, Noonan JP. Origin and evolution of developmental enhancers in the mammalian neocortex. Proc Natl Acad Sci U S A. 2016;113(19):E2617-26. Epub 2016/04/27. doi: 10.1073/pnas.1603718113. PubMed PMID: 27114548; PubMed Central PMCID: PMCPMC4868431.
15. Sebe-Pedros A, Chomsky E, Pang K, Lara-Astiaso D, Gaiti F, Mukamel Z, et al. Early metazoan cell type diversity and the evolution of multicellular gene regulation. Nat Ecol Evol. 2018;2(7):1176-88. Epub 2018/06/27. doi: 10.1038/s41559-018-0575-6. PubMed PMID: 29942020; PubMed Central PMCID: PMCPMC6040636.

16. Teichmann SA, Babu MM. Gene regulatory network growth by duplication. Nature Genetics. 2004;36(5):492-6. doi: 10.1038/ng1340.

17. Kaessmann H. Origins, evolution, and phenotypic impact of new genes. Genome Res. 2010;20(10):1313-26. Epub 2010/07/24. doi: 10.1101/gr.101386.109. PubMed PMID: 20651121; PubMed Central PMCID: PMCPMC2945180.

18. McLysaght A, Hurst LD. Open questions in the study of de novo genes: what, how and why. Nat Rev Genet. 2016;17(9):567-78. doi: 10.1038/nrg.2016.78. PubMed PMID: WOS:000381510700013.

19. Xie C, Bekpen C, Künzel S, Keshavarz M, Krebs-Wheaton R, Skrabar N, et al. 2019. doi: 10.1101/510214.

20. Betran E, Reinhardt JA, Wanjiru BM, Brant AT, Saelao P, Begun DJ, et al. De Novo ORFs in Drosophila Are Important to Organismal Fitness and Evolved Rapidly from Previously Non-coding Sequences. PLoS Genetics. 2013;9(10). doi: 10.1371/journal.pgen.1003860.

21. Kim SK, Mayer MG, Rödelsperger C, Witte H, Riebesell M, Sommer RJ. The Orphan Gene dauerless Regulates Dauer Development and Intraspecific Competition in Nematodes by Copy Number Variation. PLOS Genetics. 2015;11(6). doi: 10.1371/journal.pgen.1005146.

22. Li D, Yan Z, Lu L, Jiang H, Wang W. Pleiotropy of the de novo-originated gene MDF1. Scientific Reports. 2014;4(1). doi: 10.1038/srep07280.

23. Neme R, Tautz D. Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to de novo gene emergence. eLife. 2016;5. doi: 10.7554/eLife.09977.

24. Kapranov P, Willingham AT, Gingeras TR. Genome-wide transcription and the implications for genomic organization. Nat Rev Genet. 2007;8(6):413-23. doi: 10.1038/nrg2083.

25. Ruiz-Orera J, Verdaguer-Grau P, Villanueva-Canas JL, Messeguer X, Alba MM. Translation of neutrally evolving peptides provides a basis for de novo gene evolution. Nat Ecol Evol. 2018;2(5):890-6. Epub 2018/03/21. doi: 10.1038/s41559-018-0506-6. PubMed PMID: 29556078.

26. Ingolia Nicholas T, Brar Gloria A, Stern-Ginossar N, Harris Michael S, Talhouarne Gaëlle JS, Jackson Sarah E, et al. Ribosome Profiling Reveals Pervasive Translation Outside of Annotated Protein-Coding Genes. Cell Reports. 2014;8(5):1365-79. doi: 10.1016/j.celrep.2014.07.045.

27. Prabh N, Rödelsperger C. Are orphan genes protein-coding, prediction artifacts, or non-coding RNAs? BMC Bioinformatics. 2016;17(1). doi: 10.1186/s12859-016-1102-x.

28. Zhang L, Ren Y, Yang T, Li G, Chen J, Gschwend AR, et al. Rapid evolution of protein diversity by de novo origination in Oryza. Nature Ecology & Evolution. 2019;3(4):679-90. doi: 10.1038/s41559-019-0822-5.

29. Schmitz JF, Ullrich KK, Bornberg-Bauer E. Incipient de novo genes can evolve from frozen accidents that escaped rapid transcript turnover. Nat Ecol Evol. 2018;2(10):1626-32. Epub 2018/09/12. doi: 10.1038/s41559-018-0639-7. PubMed PMID: 30201962.

30. Ruiz-Orera J, Alba MM. Translation of Small Open Reading Frames: Roles in Regulation and Evolutionary Innovation. Trends Genet. 2019;35(3):186-98. Epub 2019/01/05. doi: 10.1016/j.tig.2018.12.003. PubMed PMID: 30606460.

31. Tautz D, Domazet-Loso T. The evolutionary origin of orphan genes. Nat Rev Genet. 2011;12(10):692-702. Epub 2011/09/01. doi: 10.1038/nrg3053. PubMed PMID: 21878963.

32. Wu X, Sharp Phillip A. Divergent Transcription: A Driving Force for New Gene Origination? Cell. 2013;155(5):990-6. doi: 10.1016/j.cell.2013.10.048.

33. Levine M, Cattoglio C, Tjian R. Looping back to leap forward: transcription enters a new era. Cell. 2014;157(1):13-25. Epub 2014/04/01. doi: 10.1016/j.cell.2014.02.009. PubMed PMID: 24679523; PubMed Central PMCID: PMCPMC4059561.

34. Kim T-K, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, et al. Widespread transcription at neuronal activity-regulated enhancers. Nature. 2010;465(7295):182-7. doi: 10.1038/nature09033.

35. Evdokimova V, Ruzanov P, Imataka H, Raught B, Svitkin Y, Ovchinnikov LP, et al. The major mRNA-associated protein YB-1 is a potent 5' cap-dependent mRNA stabilizer. EMBO J. 2001;20(19):5491-502. Epub 2001/09/28. doi: 10.1093/emboj/20.19.5491. PubMed PMID: 11574481; PubMed Central PMCID: PMCPMC125650.

36. Shatkin A. Capping of eucaryotic mRNAs. Cell. 1976;9(4):645-53. doi: 10.1016/0092-8674(76)90128-8.

37. Visa N, Izaurralde E, Ferreira J, Daneholt B, Mattaj IW. A nuclear cap-binding complex binds Balbiani ring pre-mRNA cotranscriptionally and accompanies the ribonucleoprotein particle during nuclear export. J Cell Biol. 1996;133(1):5-14. Epub 1996/04/01. PubMed PMID: 8601613; PubMed Central PMCID: PMCPMC2120770.

38. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet. 2007;39(3):311-8. Epub 2007/02/06. doi: 10.1038/ng1966. PubMed PMID: 17277777.

39. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57-74. doi: 10.1038/nature11247.

40. Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proceedings of the National Academy of Sciences. 2010;107(50):21931-6. doi: 10.1073/pnas.1016071107.

41. Berthelot C, Villar D, Horvath JE, Odom DT, Flicek P. Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. Nat Ecol Evol. 2018;2(1):152-63. Epub 2017/11/29. doi: 10.1038/s41559-017-0377-2. PubMed PMID: 29180706; PubMed Central PMCID: PMCPMC5733139.

42. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014;507(7493):455-61. doi: 10.1038/nature12787.

43. Carelli FN, Liechti A, Halbert J, Warnefors M, Kaessmann H. Repurposing of promoters and enhancers during mammalian evolution. Nat Commun. 2018;9(1):4066. Epub 2018/10/06. doi: 10.1038/s41467-018-06544-z. PubMed PMID: 30287902; PubMed Central PMCID: PMCPMC6172195.

44. Ong C-T, Corces VG. CTCF: an architectural protein bridging genome topology and function. Nat Rev Genet. 2014;15(4):234-46. doi: 10.1038/nrg3663.

45. Cusanovich DA, Hill AJ, Aghamirzaie D, Daza RM, Pliner HA, Berletch JB, et al. A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. Cell. 2018;174(5):1309-24 e18. Epub 2018/08/07. doi: 10.1016/j.cell.2018.06.052. PubMed PMID: 30078704; PubMed Central PMCID: PMCPMC6158300.

46. Neme R, Tautz D. Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. BMC Genomics. 2013;14(1). doi: 10.1186/1471-2164-14-117.

47. Tabula Muris C, Overall c, Logistical c, Organ c, processing, Library p, et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature. 2018;562(7727):367-72. Epub 2018/10/05. doi: 10.1038/s41586-018-0590-4. PubMed PMID: 30283141.

48. Kryuchkova-Mostacci N, Robinson-Rechavi M. Tissue-Specific Evolution of Protein Coding Genes in Human and Mouse. PLoS One. 2015;10(6):e0131673. Epub 2015/06/30. doi: 10.1371/journal.pone.0131673. PubMed PMID: 26121354; PubMed Central PMCID: PMCPMC4488272.

49. Mantsoki A, Devailly G, Joshi A. Gene expression variability in mammalian embryonic stem cells using single cell RNA-seq data. Computational Biology and Chemistry. 2016;63:52-61. doi: 10.1016/j.compbiolchem.2016.02.004.

50. Werner MS, Sieriebriennikov B, Prabh N, Loschko T, Lanz C, Sommer RJ. Young genes have distinct gene structure, epigenetic profiles, and transcriptional regulation. Genome Research. 2018;28(11):1675-87. doi: 10.1101/gr.234872.118.

51. Sebe-Pedros A, Saudemont B, Chomsky E, Plessier F, Mailhe MP, Renno J, et al. Cnidarian Cell Type Diversity and Regulation Revealed by Whole-Organism Single-Cell RNA-Seq. Cell. 2018;173(6):1520-34 e20. Epub 2018/06/02. doi: 10.1016/j.cell.2018.05.019. PubMed PMID: 29856957.

52. Schwaiger M, Schonauer A, Rendeiro AF, Pribitzer C, Schauer A, Gilles AF, et al. Evolutionary conservation of the eumetazoan gene regulatory landscape. Genome Research. 2014;24(4):639-50. doi: 10.1101/gr.162529.113.

53. Kvon EZ, Kazmar T, Stampfel G, Yáñez-Cuna JO, Pagani M, Schernhuber K, et al. Genome-scale functional characterization of Drosophila developmental enhancers in vivo. Nature. 2014;512(7512):91-5. doi: 10.1038/nature13395.

54. Capra JA, Pollard KS, Singh M. Novel genes exhibit distinct patterns of function acquisition and network integration. Genome Biol. 2010;11(12):R127. Epub 2010/12/29. doi: 10.1186/gb-2010-11-12-r127. PubMed PMID: 21187012; PubMed Central PMCID: PMCPMC3046487.

55. Abrusán G. Integration of New Genes into Cellular Networks, and Their Structural Maturation. Genetics. 2013;195(4):1407-17. doi: 10.1534/genetics.113.152256.

56. Warnefors M, Eyre-Walker A. The Accumulation of Gene Regulation Through Time. Genome Biology and Evolution. 2011;3:667-73. doi: 10.1093/gbe/evr019.

57. Carroll SB. Chance and necessity: the evolution of morphological complexity and diversity. Nature. 2001;409(6823):1102-9. Epub 2001/03/10. doi: 10.1038/35059227. PubMed PMID: 11234024.

58. Domazet-Lošo T, Brajković J, Tautz D. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. Trends in Genetics. 2007;23(11):533-9. doi: 10.1016/j.tig.2007.08.014.

59. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nature Methods. 2013;10(12):1213-8. doi: 10.1038/nmeth.2688.

60. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The Human Genome Browser at UCSC. Genome Research. 2002;12(6):996-1006. doi: 10.1101/gr.229102.

61. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841-2. doi: 10.1093/bioinformatics/btq033.

62. Lizio M, Harshbarger J, Shimoji H, Severin J, Kasukawa T, Sahin S, et al. Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome Biology. 2015;16(1). doi: 10.1186/s13059-014-0560-6.

63. Noguchi S, Arakawa T, Fukuda S, Furuno M, Hasegawa A, Hori F, et al. FANTOM5 CAGE profiles of human and mouse samples. Scientific Data. 2017;4. doi: 10.1038/sdata.2017.112.

64. Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proceedings of the National Academy of Sciences. 2003;100(26):15776-81. doi: 10.1073/pnas.2136655100.

65. Ingolia NT. Ribosome profiling: new views of translation, from single codons to genome scale. Nat Rev Genet. 2014;15(3):205-13. doi: 10.1038/nrg3645.

66. Michel AM, Fox G, M. Kiran A, De Bo C, O'Connor PBF, Heaphy SM, et al. GWIPS-viz: development of a ribo-seq genome browser. Nucleic Acids Research. 2014;42(D1):D859-D64. doi: 10.1093/nar/gkt1035.

67. Dalby M, Rennie, Sarah, & Andersson, Robin. FANTOM5 transcribed enhancers in mm10 Zenodo2018.

68. Pliner HA, Packer JS, McFaline-Figueroa JL, Cusanovich DA, Daza RM, Aghamirzaie D, et al. Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. Molecular Cell. 2018;71(5):858-71.e8. doi: 10.1016/j.molcel.2018.06.044.

69. Cunningham F, Achuthan P, Akanni W, Allen J, Amode MR, Armean IM, et al. Ensembl 2019. Nucleic Acids Res. 2019;47(D1):D745-D51. Epub 2018/11/09. doi: 10.1093/nar/gky1113. PubMed PMID: 30407521; PubMed Central PMCID: PMCPMC6323964.