Enhancers facilitate the birth of *de novo* genes and their 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 are adaptive. Here, we use whole-tissue and 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 intergenic open reading frames are preferentially located near enhancers, whereas older open reading frames are not. Mouse-specific intergenic open reading frames that are proximal to enhancers are more highly and stably transcribed than those that are not proximal to enhancers or promoters, and they are transcribed in a limited diversity of cellular contexts. Furthermore, we report several instances of mouse-specific intergenic open reading frames that are proximal to promoters that show evidence of being repurposed enhancers. We also show that open reading frames gradually acquire specific interactions with enhancers 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 (Catarino and Stark 2018; Haberle and Stark 2018), 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 (Davidson and Levine 2008; Spitz and Furlong 2012).

Enhancer activity is not only dynamic throughout development, but also throughout evolution (Villar, et al. 2015). 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 (Prud'homme, et al. 2007; Carroll 2008). 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 (Peter and Davidson 2011). Examples include the archetypical pentadactyl limb anatomy of extant tetrapods (Kherdjemil, et al. 2016), ocular regression in subterranean rodents (Partha, et al. 2017; Roscito, et al. 2018), limb loss in snakes (Kvon, et al. 2016; Roscito, et al. 2018), convergent pigmentation patterns in East African cichlids (Kratochwil, et al. 2018), the diversity of butterfly wing patterns (Barton, et al. 2016), the mammalian neocortex (Emera, et al. 2016), and cell type diversity in eumetazoans (Sebé-Pedrós, et al. 2018).

Regulatory networks not only evolve via rewiring, but also via the addition of new genes (Teichmann and Babu 2004). 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 (Kaessmann 2010), 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 (Carvunis, et al. 2012; Betran, et al. 2013; Li, et al. 2014; Kim, et al. 2015; McLysaght and Hurst 2016; Van Oss and Carvunis 2019; Willemsen, et al. 2019). 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 (Kapranov, et al. 2007; Neme and Tautz 2016) and many lineage-specific transcripts containing ORFs show evidence of translation (Wilson and Masel 2011; Ingolia, et al. 2014; Ruiz-Orera, et al. 2014; Prabh and Rödelsperger 2016; Ruiz-Orera, et al. 2018; Schmitz, et al. 2018; Ruiz-Orera and Alba 2019; Zhang, et al. 2019), 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 (Tautz and Domazet-Loso 2011). 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 (Wu and Sharp 2013), and thus in the growth of gene regulatory networks. The physical proximity between active enhancers and their target genes (Levine, et al. 2014) – facilitated by DNA looping – creates a transcriptionally permissive environment that is engaged with RNA polymerase II, which may lead to the transcription of DNA near the enhancer, or to the transcription of the enhancer itself, producing so-called enhancer RNA (De Santa, et al. 2010; Kim, et al. 2010; Li, et al. 2016; Haberle and Stark 2018). If the transcript contains an open reading frame, then such increased transcription will increase the likelihood of interaction with ribosomes, and because enhancers are typically active in a small number of cell types, interactions with ribosomes will occur in a limited diversity of cellular contexts. This may help purge toxic peptides and enrich for benign peptides, a process that has been hypothesized to increase the likelihood of *de novo* gene birth (Wilson and Masel 2011). Moreover, similarities in the architectures of enhancers and promoters may facilitate the regulatory repurposing of the former into the latter (Carelli, et al. 2018), reinforcing the transcription of new open reading frames that emerge near enhancers. 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, 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.

The first evidence that enhancers can facilitate *de novo* gene birth was recently provided using whole-animal transcriptomic and epigenetic data from the nematode *Pristionchus pacificus* (Werner, et al. 2018). Specifically, the transcription start sites of expressed genes that were in open chromatin and private to *P. pacificus* were found to be in closer proximity to histone modifications indicative of enhancers than the transcription start sites of expressed genes that were in open chromatin and shared with other nematode species. While this evidence is compelling, additional systematic analyses are required to draw firm conclusions and to address remaining open questions. For example, we do not yet know about the generality of this mechanism, specifically whether it applies to other clades of eumetazoa. Furthermore, information on the stability of the transcribed ORFs or their potential for translation is still lacking. We also do not know about the cell-type specificity of the enhancers that facilitate *de novo* gene birth (because the data used to study *P. pacificus* were derived from the whole animal) or how the facilitating role of enhancers in *de novo* gene birth differs from that of other means of pervasive transcription (Neme and Tautz 2016). Finally, we do not know how enhancers integrate new genes into existing cellular networks, especially over macro-evolutionary timescales.

Here, we take an integrative approach to address these open questions and to study the potential dual role of enhancers in the evolution of gene regulatory networks. We leverage whole-tissue and 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 young ORFs are preferentially located near enhancers, whereas older ORFs are not. Some of these young ORFs likely are enhancers, as evidenced by their bidirectional transcription – a hallmark of enhancer activity. Mouse-specific intergenic ORFs that are proximal to enhancers are more highly and stably transcribed than mouse-specific intergenic ORFs that are not proximal to enhancers or promoters, and they are transcribed in a limited diversity of cellular contexts, thus highlighting fundamental differences between the facilitating role of enhancers versus other forms of pervasive transcription in *de novo* gene birth. We find the transcripts of enhancer-proximal ORFs often associate with ribosomes, and we uncover several instances of mouse-specific intergenic ORFs that are proximal to promoters that are likely repurposed enhancers. Finally, we show the number of enhancer interactions per ORF increases with ORF age, which correlates with an increase in expression breadth, 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 integration into gene regulatory networks.

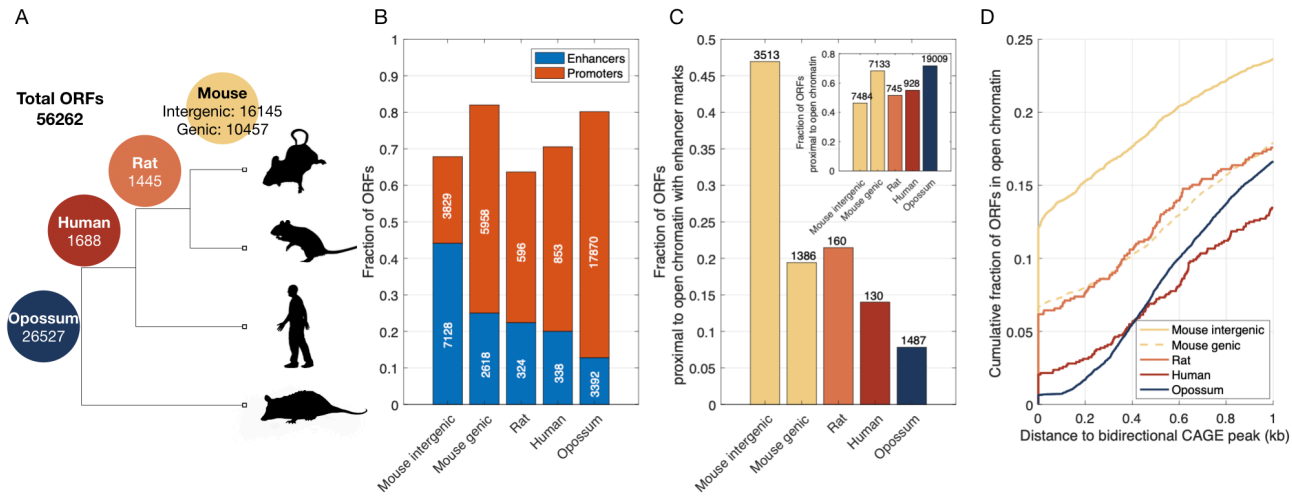
Results

Mouse-specific intergenic ORFs are often proximal to enhancers

We considered a set of 56,262 ORFs from transcripts expressed in the liver, brain, and testis of *Mus musculus.* Previous work assigned phylogenetic ages to these ORFs (Schmitz, et al. 2018), based on the presence of homologous sequences in the transcriptomes of other mammalian species, including rat, human, and opossum (Fig. 1A). We further classified the mouse-specific ORFs as genic or intergenic, based on whether or not they are proximal to older, annotated genes (Materials and Methods). We use the term *proximal* to mean within 500bp (in the supplementary material, we show our findings are qualitatively insensitive to changing this definition to 250bp and 1000bp, see Figs. S4-S8), and we use an ORF's first exon to calculate its distance from other genomic features. To characterize the regulatory background of an ORF, we considered data describing histone modifications that are indicative of promoters and enhancers (Heintzman, et al. 2007). Specifically, we merged chromatin immunoprecipitation followed by DNA sequencing (ChIP-seq) data for H3K27ac, H3K4me1, and H3K4me3 obtained from 23 mouse tissues and cell types (ENCODE 2012). We considered enhancers to be those genomic regions where H3K27ac and/or H3K4me1 peaks do not overlap H3K4me3 peaks in any tissue, and promoters to be those genomic regions with H3K4me3 peaks (Creyghton, et al. 2010; Berthelot, et al. 2018) (Materials and Methods).

The majority of ORFs in each age class are proximal to a promoter or an enhancer (Fig. 1B). Remarkably, mouse-specific intergenic ORFs are the only class of ORFs that are more likely to be proximal to enhancers than to promoters. While the first exon of nearly 45% (7,128) of mouse-specific intergenic ORFs are proximal to enhancers, fewer than 25% of rat, human, and opossum-shared ORFs are proximal to enhancers. Similar trends are observed when we restrict our attention to ORFs that are

135 within, or proximal to, genomic regions of open chromatin in at least one of thirteen mouse tissues (Fig.
136 1C; Materials and Methods). Specifically, ~47% (3,513 out of 7,484) of mouse-specific intergenic
137 ORFs are proximal to regions of open chromatin that harbor histone modifications indicative of
138 enhancers, but not promoters, whereas fewer than 21% of rat, human, and opossum-shared ORFs are
139 proximal to such regions. Similar trends are also observed when we consider histone modification data
140 from individual tissues, as opposed to merging data across cell and tissue types. Specifically, 25% (281
141 ORFs), ~36% (897 ORFs), and ~20% (537 ORFs) of intergenic mouse-specific ORFs that are in open
142 chromatin and expressed in liver, brain, and testis, respectively, are proximal to an enhancer in that
143 tissue, as compared to <10% of genic and older ORFs, which are instead preferentially proximal to
144 promoters (Fig. S2D-F). Finally, mouse-specific intergenic ORFs are more likely to show evidence of
145 bidirectional transcription – a hallmark of enhancer activity (Andersson, et al. 2014) – than any other
146 class of ORFs (Fig. 1D), with 12% of the ORFs overlapping a bidirectional CAGE peak and nearly
147 20% (1,429) of the ORFs proximal to a bidirectional CAGE peak (Fig. S1 shows that these trends are
148 not driven by exon length). Taken together, these results support a model in which enhancers facilitate
149 the expression of young ORFs (Wu and Sharp 2013; Werner, et al. 2018).
150



151

152 **Figure 1. Mouse-specific intergenic ORFs are often proximal to enhancers.** A) Phylogeny showing
153 the four age classes of the 56,262 ORFs. The numbers on the branches indicate the number of ORFs
154 that are either mouse-specific or shared with rat, human, and opossum. Mouse-specific ORFs are further
155 classified as intergenic or genic. B) Fraction of ORFs that are proximal to ChIP-seq peaks indicative of
156 enhancers (H3K27ac and/or H3K4me1 without overlapping H3K4me3) or promoters (H3K4me3),
157 shown in relation to ORF class. C) Fraction of ORFs that are proximal to regions of open chromatin
158 that contain enhancers, but not promoters, shown in relation to ORF class. The inset shows the fraction
159 of ORFs that are proximal to regions of open chromatin, regardless of whether those regions contain
160 promoters or enhancers. D) Cumulative fraction of ORFs that are proximal to regions of open
161 chromatin, shown in relation to their distance to the closest bidirectional CAGE peak.

Intergenic ORFs that are proximal to enhancers are highly and stably transcribed, relative to intergenic ORFs that are not proximal to enhancers or promoters

We next asked what differentiates the facilitating role of enhancers in *de novo* gene birth from other forms of pervasive transcription taking place away from promoters and enhancers. We hypothesized that because enhancers are regularly engaged with the transcriptional machinery, they may confer higher levels of expression and greater expression stability. To test this hypothesis, we compared the expression levels and stabilities of intergenic mouse-specific ORFs that are proximal to enhancers with those of intergenic mouse-specific ORFs that are not proximal to enhancers or promoters, using transcriptomic, histone modification, and chromatin accessibility data from liver, brain, and testis (Materials and Methods).

In all three tissues, we observed that mouse-specific intergenic ORFs that are proximal to enhancers have a higher median expression level than mouse-specific intergenic ORFs that are not proximal to enhancers or promoters (Fig. 2A-C; Wilcoxon signed-rank test, $p < 0.001$ in liver, $p = 0.003$ in brain, $p = 0.02$ in testis). To measure expression stability, we calculated the entropy of expression across biological replicates (Materials and Methods). When this measure equals its minimum of zero, the ORF is expressed in only one of the replicates; when it equals its maximum of one, the ORF is expressed at equal levels across replicates. In all three tissues, we observed that expression stability is higher for mouse-specific intergenic ORFs that are proximal to enhancers than for mouse-specific intergenic ORFs that are not proximal to enhancers or promoters (Fig. 2D-F, Wilcoxon's signed-rank test, $p < 0.001$). These observations support the hypothesis that enhancers confer higher expression levels and greater expression stability to proximal ORFs than do other forms of pervasive transcription away from promoters and enhancers.

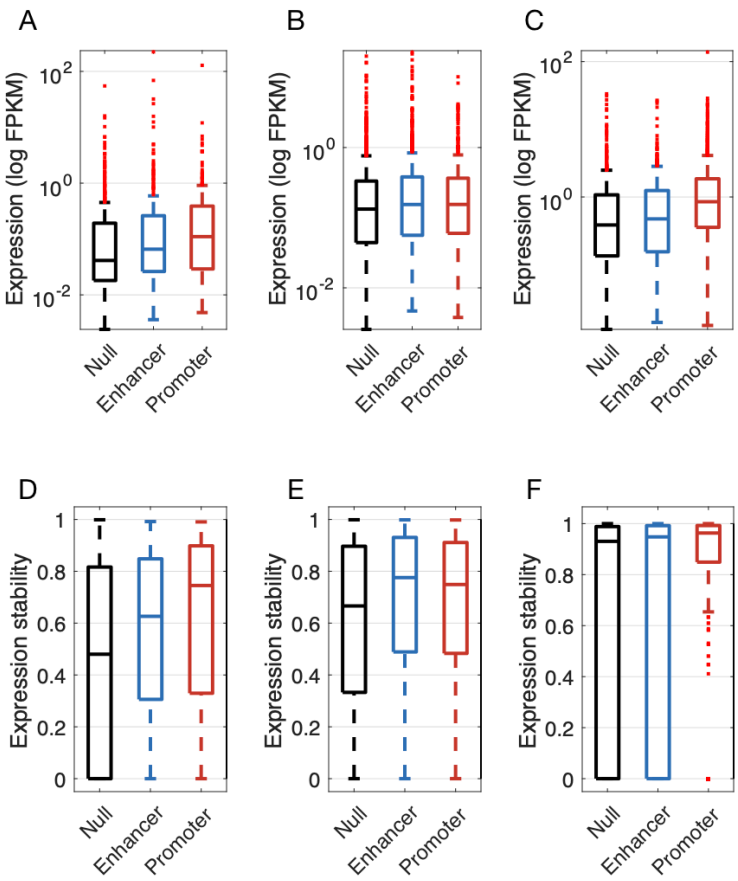


Figure 2. Mouse-specific intergenic ORFs that are proximal to enhancers are more highly expressed and have greater expression stability than mouse-specific intergenic ORFs that are not proximal to enhancers or promoters. Expression level of mouse-specific intergenic ORFs proximal to enhancers, promoters, or neither ('Null') in A) liver, B) brain, and C) testis. Expression stability of mouse-specific intergenic ORFs proximal to enhancers, promoters, or neither ('Null') in D) liver (8 replicates), E) brain (8 replicates), and F) testis (2 replicates).

Intergenic ORFs that are proximal to enhancers are more likely to associate with ribosomes than intergenic ORFs that are not proximal to enhancers or promoters

Many non-coding transcripts associate with ribosomes (Wilson and Masel 2011; Ingolia, et al. 2014; Ruiz-Orera, et al. 2014; Zhang, et al. 2015). It has been suggested that this may enrich the pool of transcribed ORFs for benign peptides, thus increasing the likelihood of *de novo* gene birth (Wilson and Masel 2011). We hypothesized that because of their increased expression levels and stabilities, mouse-specific intergenic ORFs that are proximal to enhancers will be more likely to associate with ribosomes than mouse-specific intergenic ORFs that are not proximal to enhancers or promoters. To test this hypothesis, we considered liver, brain, and testis data from a ribosomal profiling assay called ribo-seq, which describes the transcriptome-wide binding patterns of ribosomes to RNA molecules (Ruiz-Orera, et al. 2018; Ruiz-Orera and Alba 2019) (Materials and Methods).

Following Schmitz et al. (2018), we first consider a permissive definition of ribosomal association: at least one read mapping to the first exon of an ORF. We found that mouse-specific intergenic ORFs that are proximal to enhancers are ~10% more likely to associate with ribosomes than mouse-specific intergenic ORFs that are not proximal to enhancers or promoters, and ~10% less likely to associate with ribosomes than mouse-specific intergenic ORFs that are proximal to promoters (Fig. 3A). When we apply more conservative thresholds for ribosomal association, mouse-specific intergenic ORFs that are proximal to enhancers remain more likely to associate with ribosomes than mouse-specific intergenic ORFs that are not proximal to enhancers or promoters, and less likely than mouse-specific intergenic ORFs that are proximal to promoters, although the differences in ribosomal association between these classes decreases as the threshold for ribosomal association increases, both when evaluating reads per kilobase mapped to the first exon (Fig. 3A), or simply total number of reads mapped to the first exon (Fig. 3B). These trends remain when considering tissue-specific transcriptomic, histone modification, and ribosomal association data for liver, brain, and testis (Fig. 3C).

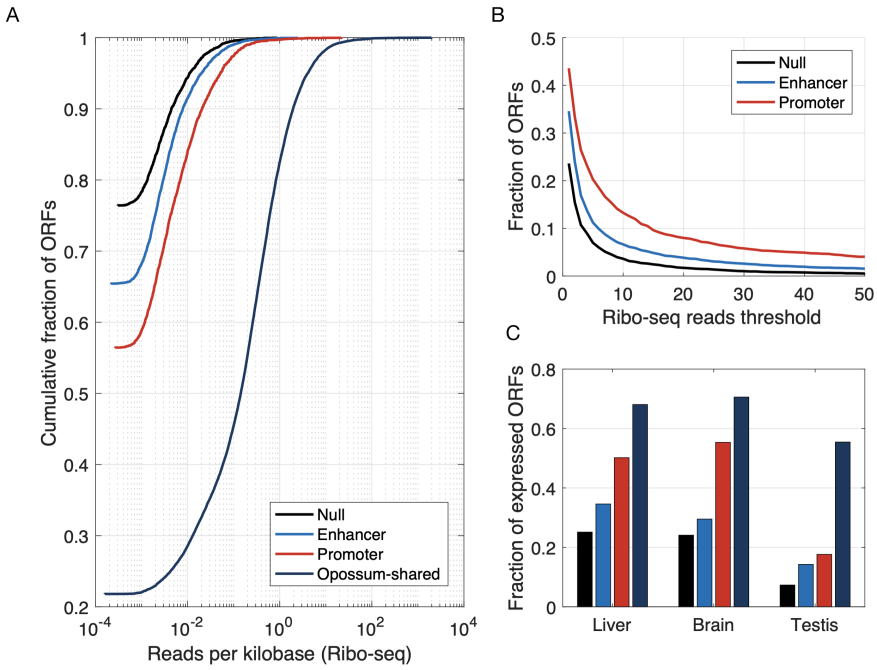
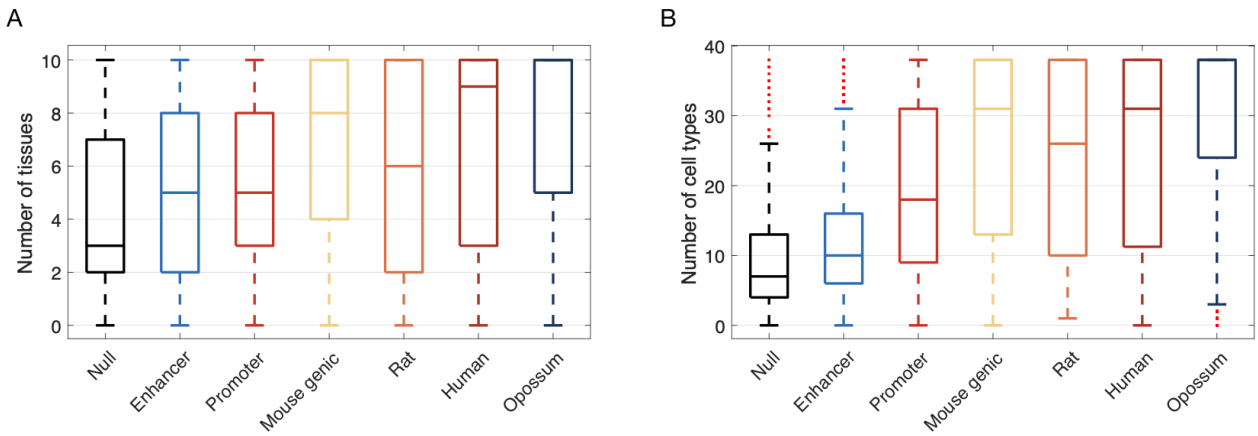


Figure 3. Mouse-specific intergenic ORFs that are proximal to enhancers are more likely to associate with ribosomes than mouse-specific intergenic ORFs that are not proximal to enhancers or promoters. A) Cumulative fractions of mouse-specific intergenic ORFs that are proximal to an enhancer, a promoter, or neither ('Null'), or to opossum-shared ORFs, shown in relation to the number of ribo-seq reads mapped per kilobase to the first exon of each ORF. B) Fraction of ORFs with ribosomal association, shown in relation to the minimum threshold for the number of reads mapped. C) Fraction of ORFs expressed in liver, brain, and testis for which at least one tissue-specific ribo-seq read could be mapped to their first exon.

227 ***Intergenic ORFs that are proximal to enhancers are expressed in more cellular contexts than***
228 ***intergenic ORFs that are not proximal to enhancers or promoters***

229 In the model of enhancer-facilitated *de novo* gene birth studied here, ORFs emerging near
230 enhancers are likely to have their expression restricted to cells where those enhancers are active.
231 Enhancers are often specific to a small number of cell types (He, et al. 2014), which may reduce the
232 potential for enhancer-proximal ORFs to have deleterious pleiotropic effects, while simultaneously
233 exposing the ORFs to a limited diversity of cellular contexts in which they may confer a selective
234 advantage. To study the breadth of expression of ORFs, we considered two sources of data: whole tissue
235 measurements of total RNA across 10 tissues, and single-cell measurements of open chromatin across
236 38 cell types (Materials and Methods).

237 We found that mouse-specific intergenic ORFs that are proximal to enhancers are expressed in
238 more tissues (Wilcoxon's signed rank test, $p < 0.001$; Fig. 4A) and are in open chromatin in more cell
239 types (Wilcoxon's signed rank test, $p < 0.001$; Fig. 4B) than mouse-specific intergenic ORFs that are
240 not proximal to enhancers or promoters. However, these ORFs are expressed in fewer tissues
241 (Wilcoxon's signed rank test, $p < 0.001$; Fig. 4A) and are in open chromatin in fewer cell types
242 (Wilcoxon's signed rank test, $p < 0.001$; Fig. 4B) than mouse-specific intergenic ORFs that are proximal
243 to promoters, genic ORFs, and ORFs shared with rat, human, and opossum. ORFs emerging near
244 enhancers are therefore transcribed in a limited number of cellular contexts, providing opportunities to
245 purge toxic peptides, and potentially balancing the risk of deleterious pleiotropic effects with the reward
246 of sampling a diversity of cellular environments.
247



248

249 **Figure 4. Mouse-specific intergenic ORFs that are proximal to enhancers are expressed in a**
250 **limited diversity of cellular contexts.** A) Number of tissues in which ORFs have an average FPKM >
251 0 across replicates. B) Number of cell types in which ORFs are in regions of open chromatin. In both
252 panels, the "Null", "Enhancer", and "Promoter" categories correspond to mouse-specific intergenic
253 ORFs.

254 ***Some intergenic ORFs are proximal to promoters that show evidence of being repurposed enhancers***

255 Similarities in the architectures of enhancers and promoters can facilitate the regulatory
256 repurposing of enhancers into promoters (Wu and Sharp 2013; Carelli, et al. 2018), which could
257 reinforce the transcription of ORFs emerging near enhancers. We next assessed whether the mouse-
258 specific intergenic ORFs that are proximal to promoters are cases of ORFs transcribed from enhancers
259 that were repurposed into promoters. To do so, we considered 329 mouse-specific intergenic ORFs that
260 are expressed in mouse liver and proximal to an active promoter in mouse liver and that could be
261 mapped to the rat genome via liftOver and (Fig. S2D; Materials and Methods). Subsequently, we
262 assessed the chromatin modification status in the rat liver of those mapped genomic regions, using
263 ChIP-seq data for H3K27ac and H3K4me3, marking enhancers and promoters, respectively.

264 Of the regions mapped to the rat genome, 256 are proximal to H3K27ac peaks and 183 are
265 proximal to H3K4me3 peaks. The majority (70%) of the regions that are proximal to H3K27ac peaks
266 are also proximal to H3K4me3 peaks (Fig. 5A,B), implying they act as promoters in the liver of both
267 mouse and rat. However, some mapped genomic regions are at such distances from H3K4me3 peaks
268 that they could well be enhancers in rat and may therefore have been repurposed into promoters on the
269 lineage to mouse (Fig. 5B). Considering those mapped genomic regions with H3K27ac peaks that are
270 separated from an H3K4me3 peak by a conservative threshold of at least 5kb, we found 33 candidates
271 for the repurposing of rat enhancers to mouse promoters (~10% of the 329 ORFs; Fig. 5B). The ORFs
272 corresponding to these mapped genomic regions show evidence of stable transcription, both in terms of
273 expression stability across biological replicates (Fig. 5C) and in terms of their proximity to capped
274 analysis of gene expression (CAGE) peaks (Fig. 5D), which provides evidence that the transcripts are
275 5'-capped. Of note, many of these CAGE peaks are bidirectional, despite their proximity to H3K4me3
276 (promoter) peaks, which further supports the hypothesis of an enhancer origin (Fig. 5D). Finally, 79%
277 of the 33 ORFs show evidence of association with ribosomes in mouse (using our most permissive
278 criterion), which is more than we would expect by randomly sampling mouse-specific intergenic ORFs
279 that are proximal to promoters and expressed in liver (Fig. 3C; binomial test $p = 0.013$). Together, these
280 observations give further support to a model in which enhancers provide fertile ground for *de novo* gene
281 birth.

282 An alternative interpretation of these data is that promoters were repurposed as enhancers on
283 the rat lineage, rather than enhancers being repurposed as promoters on the mouse lineage. To study the
284 directionality of the repurposing, we considered ChIP-seq data for H3K27ac and H3K4me3 from the
285 liver of rabbit, which served as an outgroup (Villar, et al. 2015). Of the 33 candidate genomic regions,
286 9 could be mapped to the rabbit genome using liftOver and were proximal to an H3K27ac peak in rabbit
287 liver. Of these, 7 were proximal to an H3K27ac peak that was separated from an H3K4me3 peak by at
288 least 5kb (see for example fig. 5E). This provides further support for the hypothesis of an ancestral
289 enhancer state, at least for these 7 ORFs. Of note, all of the ORFs corresponding to these mapped
290 genomic regions are surrounded by other enhancer marks in the mouse liver (Fig. S3), which hints that

enhancer redundancy may help prevent conflicts that arise in the repurposing of enhancers into promoters, a possibility we revisit in the discussion.

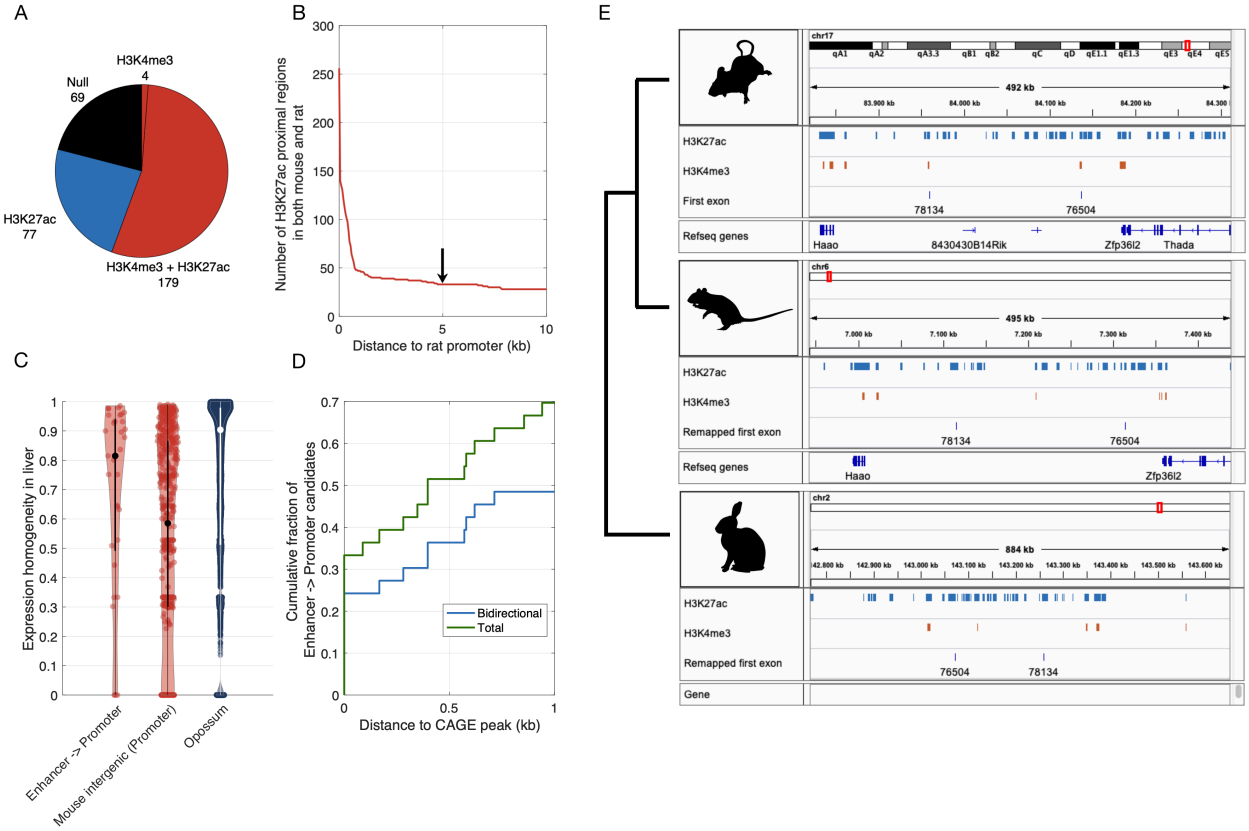


Figure 5. Some mouse-specific intergenic ORFs that are proximal to promoters show evidence of being repurposed enhancers. A) Distribution of histone methylation marks among genomic regions in rat. These regions are orthologous to genomic regions in mouse that harbour ORFs that are expressed in liver and are proximal to promoters. B) Number of genomic regions in rat liver with H3K27ac peaks, shown in relation to their distance to the closest H3K4me3 peak. We use a conservative threshold of 5kb (black arrow) between a promoter and an enhancer mark to determine that an enhancer is not a promoter. This results in 33 candidate promoters that were potentially repurposed from enhancers. C) Expression homogeneity in liver of these 33 ORFs ('Enhancer -> Promoter'), mouse-specific intergenic ORFs that are proximal to promoters in liver, and opossum-shared ORFs. D) Cumulative fraction of the 33 ORFs shown in relation to their distance to the nearest CAGE peak ('Total') or the nearest bidirectional CAGE peak ('Bidirectional'). E) Example repurposed enhancers. Orthologous genomic regions in mouse, rat, and rabbit that in mouse include the first exon of an intergenic mouse-specific ORF. The blue tracks represent H3K27ac peaks (enhancers) and the red tracks represent H3K4me3 peaks (promoters), both measured in liver samples from each organism. Annotated refseq genes are also indicated.

311 ***Enhancer interactions are gradually acquired over macro-evolutionary timescales***

312 We next asked how enhancers integrate new genes into existing regulatory networks. To do so,
313 we considered an enhancer-promoter interaction map derived from single-cell chromatin accessibility
314 data in 13 murine tissues (Cusanovich, et al. 2018) (Materials and Methods). We uncovered a positive
315 correlation between the age of an ORF and its number of enhancer interactions (Spearman's correlation
316 coefficient $\rho = 0.23$, $p < 0.001$; Fig. 6A). Among mouse-specific ORFs, intergenic ORFs had fewer
317 enhancer interactions than genic ORFs, which were similar to non-mouse-specific ORFs in their
318 number of enhancer interactions. This suggests that mouse-specific ORFs of genic origin tend to coopt
319 the regulatory interactions of their host gene, or of nearby genes. Overall, the number of enhancer
320 interactions increases from a median of 6 for mouse-specific intergenic ORFs to a median of 21 for
321 ORFs that are shared with opossum, indicating that enhancer-promoter interactions are acquired
322 gradually over time.

323 To explore the pace at which enhancer interactions are acquired over macro-evolutionary
324 timescales, we shifted our focus to opossum-shared ORFs: we considered 13,217 opossum-shared
325 ORFs corresponding to annotated genes with different first exons and separated them into 15 new age
326 classes dating back to the origin of cellular life (Neme and Tautz 2013) (Fig. 6B). We again found a
327 significant correlation between the age of a gene and its number of enhancer interactions (Spearman's
328 correlation coefficient $\rho = 0.10$, $p < 0.001$; Fig. 6C).

329 We next explored the consequences of enhancer acquisition. First, we studied the expression
330 breadth of opossum-shared annotated genes using single-cell transcriptomic data from 68 cell types of
331 ten murine tissues (Tabula Muris Consortium 2018), for which single-cell chromatin accessibility data
332 were also available (Materials and Methods). We found that expression breadth increases with the
333 number of enhancer interactions (Spearman's correlation coefficient $\rho = 0.49$, $p < 0.001$; Fig. 6D) and
334 with gene age (Spearman's correlation coefficient $\rho = 0.13$, $p < 0.001$). The latter observation
335 corroborates previous findings based on transcriptomic data from whole tissues (Kryuchkova-Mostacci
336 and Robinson-Rechavi 2015). We next quantified the expression homogeneity of each gene across all
337 cell types where expression was measurable (Materials and Methods), uncovering a positive correlation
338 between expression homogeneity and the number of enhancer interactions (Spearman's correlation
339 coefficient $\rho = 0.45$, $p < 0.001$; Fig 6E), as well as gene age (Spearman's correlation coefficient $\rho =$
340 0.11, $p < 0.001$).

341 Taken together, these results show that genes acquire enhancer interactions gradually over
342 macro-evolutionary timescales, a process that correlates with expression breadth and homogeneity
343 across cell types. Enhancers thus facilitate the integration of genes into regulatory networks.
344

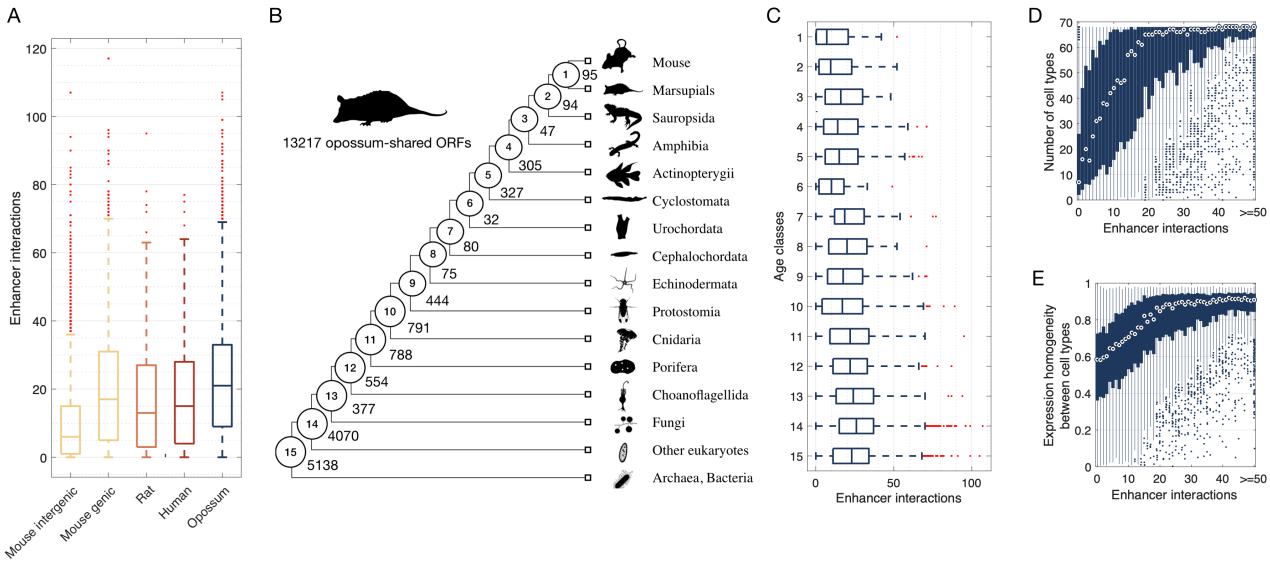


Figure 6. Enhancers facilitate the integration of genes into regulatory networks. A) Number of enhancer interactions per ORF. B) Phylogeny adapted from (Neme and Tautz 2013). The numbered circles indicate lineages representative of the age classes to which we assigned 13,217 opossum-shared ORFs, each of which corresponds to an annotated gene. The numbers on each branch represent the total number of annotated genes assigned to each age class. C) Number of enhancer interactions per gene, shown in relation to the age classes depicted in B. D) Expression breadth and E) homogeneity of opossum-shared annotated genes as a function of the number of enhancer interactions.

Discussion

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 (Kim, et al. 2010) and later expanded upon in a perspective piece by Wu and Sharp (Wu and Sharp 2013). Our findings complement recent work 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 (Werner, et al. 2018). The observation that many young ORFs are proximal to enhancers in both nematodes and mammals suggests that this mode of gene evolution dates back to at least the common ancestor of Bilateria, and possibly even earlier, since cnidarians, ctenophores, and sponges also employ distal regulatory elements (Schwaiger, et al. 2014; Gaiti, et al. 2017; Sebé-Pedrós, et al. 2018; Sebé-Pedrós, et al. 2018). As the complexity of gene regulation increased during the evolution of some lineages, such as the lineage to vertebrates (Marletaz, et al. 2018), we speculate that enhancer-facilitated *de novo* gene birth may have played an increasingly prominent role in the expansion of gene regulatory networks.

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

370 underlies the "out-of-testis hypothesis," which proposes the testis as a primary tissue for the origination
371 of new genes (Kaessmann 2010; Witt, et al. 2019). Both scenarios envision regions of open chromatin
372 that are exposed to the transcriptional machinery, and thus produce a transcriptionally active
373 environment that is conducive to the evolution of new genes. The two scenarios differ, however, in at
374 least three ways. First, genes that emerge from or near enhancers may rapidly acquire their own
375 promoters, due to the similar architectural and functional features of enhancers and promoters, a
376 similarity that facilitates the repurposing of the former to the latter (Wu and Sharp 2013). Indeed, we
377 report several mouse-specific intergenic ORFs that are proximal to promoters that show evidence of
378 being repurposed enhancers, complementing recent analyses of enhancer repurposing in primates and
379 rodents (Carelli, et al. 2018). Second, enhancers are often deployed in multiple cell types or
380 developmental stages (Kvon, et al. 2014), exposing enhancer-proximal young ORFs to selection in a
381 limited diversity of cellular contexts. This may help to purge toxic peptides (Wilson and Masel 2011),
382 and balance the benefit of expression in distinct cellular environments with the cost of pleiotropic
383 effects. Third, because enhancers are often active in somatic cell types, *de novo* genes emerging near
384 enhancers are more likely to be involved in physiological or morphological traits than *de novo* genes
385 emerging from testis, which are more likely to be involved in reproductive traits. We emphasize that
386 the enhancer-facilitated and out-of-testis scenarios are not mutually exclusive; in fact, they may be
387 complementary or even interactive. Indeed, we found many young transcribed ORFs that associate with
388 ribosomes in testis that are also proximal to enhancers (Figs. S2F, 3C).

389 The three points that differentiate enhancer-facilitated *de novo* gene birth from the out-of-testis
390 scenario also differentiate enhancer-facilitated *de novo* gene birth from pervasive transcription (Clark,
391 et al. 2011) taking place away from promoters or enhancers. An additional difference is the relatively
392 high and stable expression levels of enhancers, which increases the chances of ORF-bearing transcripts
393 that stem from or near enhancers to associate with ribosomes. This is indeed what we observe when
394 comparing ribosomal association among mouse-specific intergenic ORFs that are proximal to enhancers
395 with mouse-specific intergenic ORFs that are not proximal to enhancers or promoters. However, we
396 note that this observation may be a technological artifact. If the likelihood of the ribo-seq assay to detect
397 ribosomal association increases with the level or stability of expression, then we would expect to see
398 increased ribosomal association for mouse-specific intergenic ORFs that are proximal to enhancers,
399 relative to mouse-specific intergenic ORFs that are not proximal to enhancers or promoters, even if
400 these two classes of ORFs tend to associate with ribosomes to the same extent. Thus, the reason why
401 enhancer proximity increases the likelihood of ribosomal association is the same reason why we cannot
402 rule out the possibility that we observe this association due to a technological artifact.

403 An additional facet to enhancer-facilitated *de novo* gene birth is conflict between the enhancer and
404 the emerging gene. If the enhancer is repurposed as a promoter to enforce directional transcription, then
405 the ancestral function of the enhancer may be compromised. There are at least two ways to resolve this

406 conflict. One is to maintain enhancer function; indeed, many promoters also act as enhancers (Medina-
407 Rivera, et al. 2018). Another is enhancer redundancy. Genes are often targeted by multiple enhancers,
408 and in many of these cases, only a subset of the enhancers are necessary to drive correct expression
409 under normal growth conditions (Osterwalder, et al. 2018). Thus, we hypothesize that redundant
410 enhancers are less likely to face conflict in facilitating *de novo* gene birth. While our observation that
411 repurposed enhancers tend to be surrounded by other enhancers provides anecdotal support for this
412 hypothesis (Fig. S3), more systematic analyses are warranted.

413 The hypothesis that enhancers help *de novo* genes integrate into existing regulatory networks was
414 previously proposed in the context of the out-of-testis hypothesis, as a means to expand a new gene's
415 breadth of expression (Tautz and Domazet-Loso 2011). Using single-cell chromatin accessibility and
416 transcriptomic data, our study provides empirical support for the hypothesis that *de novo* genes
417 gradually acquire enhancer interactions over time, and that this acquisition increases expression breadth
418 and homogeneity. These findings complement related studies of gene integration into cellular networks,
419 such as networks of protein-protein interactions (Capra, et al. 2010; Abrusán 2013). Our observation
420 that genes continue to acquire enhancer interactions over macro-evolutionary timescales mirrors similar
421 increases in other aspects of gene regulation, such as in the number of proximal transcription factor
422 binding sites, alternative transcript isoforms, and miRNA targets (Warnefors and Eyre-Walker 2011).

423 Regulatory networks drive the spatiotemporal gene expression patterns that give rise to and define
424 the numerous and distinct cellular identities characteristic of Metazoan life. Enhancers play an integral
425 role in this process, mediating cell-type-specific gene-gene interactions, thus facilitating the
426 combinatorial deployment of different genes and network modules in different contexts. Genetic
427 changes that affect such interactions are responsible for myriad evolutionary adaptations and
428 innovations (Carroll 2001; Prud'homme, et al. 2007; Carroll 2008; Peter and Davidson 2011). Our
429 results suggest that the power of enhancers in creating such evolutionary novelties lies not only in their
430 ability to rewire gene regulatory networks, but also in their ability to expand them, by providing fertile
431 ground for *de novo* gene birth.

432

433 **Materials and methods**

434 *ORF age and classification*

435 Schmitz et al. (2018) identified a set of 58,864 ORFs from the transcriptomes of three murine
436 tissues: liver, brain, and testis. Blasting against the transcriptomes of four other mammalian species (rat,
437 human, kangaroo rat, and opossum), they estimated the age of each ORF by phylostratigraphic methods
438 (Domazet-Lošo, et al. 2007; Schmitz, et al. 2018). Because of the small number of ORFs shared with
439 the kangaroo rat (49 ORFs), we merged these ORFs together with those from the rat age class. We used
440 the genomic coordinates of the first exon of each ORF in the mm10 mouse genome reference to study

441 the regulatory properties of ORFs of different ages, for example to study their distance to the nearest
442 enhancer. We only considered ORFs that were transcribed from nuclear chromosomes and whose first
443 exon was longer than 30 base pairs. If first exons were shared between more than one ORF, we only
444 retained the oldest of the ORFs. Our filtered dataset contained 56,262 ORFs.
445 Schmitz et al. (2018) annotated each ORF as belonging to one of 8 different categories:
446 "intergenic," "close to promoter same strand," "close to promoter opposite strand," "overlapping same
447 strand," "overlapping opposite strand," "overlapping coding sequence same strand," "overlapping
448 coding sequence opposite strand," and "overlapping annotated gene in frame." We considered all
449 categories except "intergenic" to be "genic" in order to separate ORFs that were born within or near
450 existing genes from those that were not. This resulted in 5 classes: Mouse-specific intergenic ORFs,
451 mouse-specific genic ORFs, rat-shared ORFs, human-shared ORFs and opossum-shared ORFs.
452
453 *Proximity to enhancers and promoters*
454 We obtained ChIP-seq data for H3K27ac, H3K4me1, and H3K4me3 modifications from 23
455 different tissues and cell types from the ENCODE project (bone marrow, cerebellum, cortex, heart,
456 kidney, liver, lung, olfactory bulb, placenta, spleen, small intestine, testis, thymus, embryonic whole
457 brain, embryonic liver, embryonic limb, brown adipose tissue, macrophages, MEL, MEF, mESC, CH12
458 cell line, and E14 embryonic mouse) (ENCODE 2012). We used liftOver (Kent, et al. 2002) to convert
459 the genomic coordinates of the peaks from mm9 to mm10. We used the "merge" function of bedtools
460 (Quinlan and Hall 2010) with default parameters to collate the peaks for all tissues and cell types,
461 considering any overlapping H3K27ac and H3K4me1 peak as part of the same enhancer. We used the
462 "intersect" function of bedtools with default parameters to separate H3K27ac and H3K4me1 peaks that
463 overlapped any length of H3K4me3 peaks from those that did not. This resulted in 172,930 H3K27ac
464 and 277,187 H3K4me1 peaks that did not overlap H3K4me3 peaks. We considered genomic regions
465 with H3K4me3 peaks to be promoters, and those exclusively with H3K27ac and/or H3K4me1 peaks to
466 be enhancers (Berthelot, et al. 2018). We measured the distance in base pairs between the first exon of
467 an ORF to an enhancer or promoter using the "closest" function of bedtools with the "-t first" option
468 activated. We considered an ORF to be proximal to an enhancer if the distance to the first exon was
469 shorter than 500 bp and there was no promoter within that distance. When controlling for the length of
470 the first exon, we considered the distance to windows of 750 bp up- and downstream of the central
471 nucleotide of the first exon, rather than to the first exon itself.
472 We followed the same procedures when measuring the distance of ORFs to enhancers and
473 promoters in liver, brain, and testis tissues separately. For brain tissue, we merged ChIP-seq data from
474 embryonic whole brain and cortex. The ORFs we considered as expressed in each tissue (Fig. S2A-C)
475 were those with a mean FPKM greater than zero across replicates of total RNA transcriptomic data (8
476 replicates for liver and brain, and 2 replicates for testis) (Li, et al. 2017).
477

Chromatin accessibility

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). We obtained the data from the Mouse ATAC atlas (Cusanovich, et al. 2018), which comprised 436,206 peaks of open chromatin. We used liftOver to convert the genome coordinates from mm9 to mm10. A total of 29 peaks could not be converted. Using the "closest" function of bedtools with the "-t first" option activated, we calculated the distance between ORFs and regions of open chromatin. We annotated regions of open chromatin as enhancers if they overlapped H3K27ac and/or H3K4me1 peaks but not H3K4me3 peaks, or as promoters if they overlapped H3K4me3 peaks. To do so, we used the bedtools intersect function with the -u option activated.

Cusanovich et al. (2018) used these single-cell ATAC-seq data to identify clusters of cells with similar patterns of chromatin accessibility. They assigned the clusters to 38 distinct cell types based on the chromatin accessibility of marker genes indicative of each cell type. We used these data to identify tissues and cell types where ORFs are in accessible chromatin. We considered an ORF-containing region of the genome to be in open chromatin in a certain cell type if it was accessible in at least 1% of the cells that made up at least one of the clusters of that cell type (Cusanovich et al. 2018).

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 (Lizio, et al. 2015; Noguchi, et al. 2017). This method is based on the capture of 5' capped ends of mRNA, which allows the mapping of regions of transcription initiation genome-wide (Shiraki, et al. 2003). Using the "closest" function from bedtools with the "-t first" option activated (Quinlan and Hall 2010), we measured the distance between an ORF's first exon and its closest CAGE peak. In the same manner, we also considered a subset of CAGE peaks which were annotated as bidirectional and transcribed from enhancers (Andersson, et al. 2014; Dalby, et al. 2018).

Expression level and stability

We measured the expression levels and stabilities of ORFs. To do so, we aligned paired reads produced by RNAseq from total RNA from 10 tissues (liver, testis, brain, muscle, bone, small intestine, thymus, heart, lung and spleen) (Li, et al. 2017) using STAR 2.5.3a (Dobin, et al. 2013) to the mm10 build of the mouse genome. We chose these tissues because ChIP-seq data for histone modifications were also available. For each ORF, we calculated fragments per kilobase per million mapped reads (FPKM) as the number of reads mapped to the first exon divided by a millionth of the number of reads sequenced in each sample and then by the length of the exon in kilobases. We considered an ORF to be expressed if it had an average FPKM greater than 0 across replicates.

515 We calculated the expression stability of ORF *k* as:

516 $$H_k = -\sum_{i=1}^{n} \frac{FPKM_i}{\sum_{j=1}^{n} FPKM_j} . \log_n \left(\frac{FPKM_i}{\sum_{j=1}^{n} FPKM_j}\right)$$

517 where *n* is the number of replicates for a given tissue (8 for liver and brain, 2 for testis). We refer to this
518 measure as expression homogeneity when calculated across tissues or cell types, rather than across
519 replicates for the same tissue or cell type.

520

521 *Ribosome association*

522 We used ribosome profiling (ribo-seq) data from mouse liver, brain, and testis (Ingolia 2014).
523 We obtained the coordinates of mRNA segments detected by ribo-seq from GWIPS-viz (Michel, et al.
524 2014), a database that includes such data from different studies. From this source, we considered
525 samples from liver (3 samples from 3 studies), brain (5 samples from 2 studies), and testis (1 sample)
526 (details provided in Supplementary text 1). We combined the datasets for each tissue and merged the
527 provided genomic coordinates using the bedtools merge function; we did so with the options "-c" and
528 "-o absmax" activated. Following Ruiz-Orera et al. (2018), we removed all merged coordinates shorter
529 than 26 bp, because these could be anomalous reads. We subsequently mapped these merged
530 coordinates on the first exon of our set of ORFs using the bedtools "map" function and we summed the
531 number of reads from each of the mapped merged coordinates. In this way, we were able to assign a
532 number of ribo-seq reads to each ORF, which allowed us to estimate ribosomal association and thus
533 potential for translation.

534

535 *Enhancer repurposing*

536 We considered the set of 544 mouse-specific intergenic ORFs that were transcribed (average
537 FPKM > 0) and proximal to an H3K4me3 peak in mouse liver. We filtered this set to the 357 ORFs that
538 were proximal to what Villar et al. (2015) considered to be replicated H3K4me3 peaks in mouse, in
539 order to facilitate comparison with the histone methylation data from rat and rabbit that were generated
540 for the same study. We used liftOver to map the genomic coordinates of these ORFs to the rat and rabbit
541 genomes (builds r5 and oryCun2), requiring a minimum fraction of remapped bases of 0.6 and 0.4
542 respectively (Carelli, et al. 2018). This resulted in 329 and 130 presumably orthologous genomic
543 regions in rat and rabbit respectively. Considering H3K27ac and H3K4me3 ChIP-seq peaks in the livers
544 of mouse, rat, and rabbit (Villar, et al. 2015), we then calculated the distance between these mapped
545 regions and H3K4me3 and H3K27ac peaks using the bedtools "closest" function. We considered the
546 promoter of an ORF in mouse to show evidence of being a repurposed enhancer if its mapped genomic
547 region in rat or rabbit was proximal to an H3K27ac peak, yet more than 5 kb from an H3K4me3 peak,
548 in rat or rabbit liver.

549

550 *Enhancer interactions*

551 Cusanovich et al. (Cusanovich, et al. 2018) used single-cell ATAC-seq data to predict physical
552 interactions between regions of open chromatin (Pliner, et al. 2018), thus creating an atlas of enhancer
553 interactions in single murine cells. We downloaded these data from the Mouse ATAC atlas
554 (Cusanovich, et al. 2018), which includes the cell clusters where the interactions occur, as well as the
555 co-accessibility scores of pairs of regions of open chromatin – a measure of interaction strength. We
556 disregarded cell clusters classified as "unknown" or "collisions", as well as interactions with a co-
557 accessibility score lower than 0.25, following Pliner et al. (2018). We also filtered out interactions with
558 regions of open chromatin that overlapped ChIP-seq peaks for H3K4me3 marks or no enhancer marks,
559 in order to focus solely on interactions with enhancers. An interaction was assigned to an ORF if the
560 ORF's first exon was included in the interaction.

561

562 *Age of annotated genes*

563 To study how genes acquire enhancer interactions over macro-evolutionary timescales, we
564 considered the subset of ORFs that belong to the opossum age class in Schmitz et al. (2018) and that
565 are annotated as genes in the latest version of Ensembl (release 95) (Cunningham, et al. 2019). We
566 matched these genes to age estimates reported by Neme & Tautz (2013), based on a phylostratigraphic
567 analysis of 20 lineages spanning 4 billion years from the last universal common ancestor to the common
568 ancestor of mouse and rat. We further filtered the dataset to only include ORFs that emerged in the first
569 15 of the 20 phylostrata, in order to focus on ORFs that are considered to have emerged before the split
570 between the common ancestor of placental mammals and marsupials by both Schmitz et al. (2018) and
571 Neme & Tautz (2013). This left us with ~16,000 ORFs corresponding to 13,217 unique annotated genes
572 that emerged prior to the origin of placental mammals.

573

574 *Expression breadth and homogeneity of annotated genes*

575 To study the transcription of annotated genes, we used the expression data reported by the Tabula
576 Muris Consortium (Tabula Muris Consortium 2018) for the single-cell RNA sequencing performed
577 with FACS-based cell capture in plates, for 20 different mouse tissues. The data include the log-
578 normalization of 1 + counts per million for each of the annotated genes in each of the sequenced cells.
579 We considered ten tissues that were also used for the construction of the Mouse ATAC Atlas
580 (Cusanovich, et al. 2018). We measured the expression breadth of each ORF corresponding to an
581 annotated gene as the number of cell types in which expression could be detected in at least 1% of the
582 cells assigned to a cell type. The homogeneity of expression across cell types was calculated as
583 explained above for H_k, but considering the mean expression in each cell type where expression was
584 detectable, rather than the mean expression across replicates.

585 **References**

586 Abrusán G. 2013. Integration of new genes into cellular networks, and their structural maturation.
587 *Genetics* 195:1407-1417.
588 Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl
589 C, Suzuki T, et al. 2014. An atlas of active enhancers across human cell types and tissues. *Nature*
590 507:455-461.
591 Barton NH, Wallbank RWR, Baxter SW, Pardo-Diaz C, Hanly JJ, Martin SH, Mallet J, Dasmahapatra
592 KK, Salazar C, Joron M, et al. 2016. Evolutionary Novelty in a Butterfly Wing Pattern through
593 Enhancer Shuffling. *PLOS Biol* 14.
594 Berthelot C, Villar D, Horvath JE, Odom DT, Flicek P. 2018. Complexity and conservation of
595 regulatory landscapes underlie evolutionary resilience of mammalian gene expression. *Nat Ecol Evol*
596 2:152-163.
597 Betran E, Reinhardt JA, Wanjiru BM, Brant AT, Saelao P, Begun DJ, Jones CD. 2013. De novo
598 ORFs in Drosophila are important to organismal fitness and evolved rapidly from previously non-
599 coding sequences. *PLoS Genet* 9.
600 Capra JA, Pollard KS, Singh M. 2010. Novel genes exhibit distinct patterns of function acquisition
601 and network integration. *Genome Biol* 11:R127.
602 Carelli FN, Liechti A, Halbert J, Warnefors M, Kaessmann H. 2018. Repurposing of promoters and
603 enhancers during mammalian evolution. *Nat Commun* 9:4066.
604 Carroll SB. 2001. Chance and necessity: the evolution of morphological complexity and diversity.
605 *Nature* 409:1102-1109.
606 Carroll SB. 2008. Evo-Devo and an expanding evolutionary synthesis: A genetic theory of
607 morphological evolution. *Cell* 134:25-36.
608 Carvunis AR, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, Charloteaux B,
609 Hidalgo CA, Barbette J, Santhanam B, et al. 2012. Proto-genes and de novo gene birth. *Nature*
610 487:370-374.
611 Catarino RR, Stark A. 2018. Assessing sufficiency and necessity of enhancer activities for gene
612 expression and the mechanisms of transcription activation. *Genes Dev* 32:202-223.
613 Clark MB, Amaral PP, Schlesinger FJ, Dinger ME, Taft RJ, Rinn JL, Ponting CP, Stadler PF, Morris
614 KV, Morillon A, et al. 2011. The reality of pervasive transcription. *PLoS Biol* 9:e1000625; discussion
615 e1001102.
616 Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA,
617 Frampton GM, Sharp PA, et al. 2010. Histone H3K27ac separates active from poised enhancers and
618 predicts developmental state. *Proc Natl Acad Sci U S A* 107:21931-21936.
619 Cunningham F, Achuthan P, Akanni W, Allen J, Amode MR, Armean IM, Bennett R, Bhai J, Billis
620 K, Boddu S, et al. 2019. Ensembl 2019. *Nucleic Acids Res* 47:D745-D751.

621 Cusanovich DA, Hill AJ, Aghamirzaie D, Daza RM, Pliner HA, Berletch JB, Filippova GN, Huang
622 X, Christiansen L, DeWitt WS, et al. 2018. A single-cell atlas of in vivo mammalian chromatin
623 accessibility. *Cell* 174:1309-1324 e1318.
624 Dalby M, Rennie S, Andersson R. 2018. FANTOM5 transcribed enhancers in mm10. *Zenodo*.
625 Davidson EH, Levine MS. 2008. Properties of developmental gene regulatory networks. *Proc Natl*
626 *Acad Sci U S A* 105:20063-20066.
627 De Santa F, Barozzi I, Mietton F, Ghisletti S, Polletti S, Tusi BK, Muller H, Ragoussis J, Wei CL,
628 Natoli G. 2010. A large fraction of extragenic RNA pol II transcription sites overlap enhancers. *PLoS*
629 *Biol* 8:e1000384.
630 Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR.
631 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29:15-21.
632 Domazet-Lošo T, Brajković J, Tautz D. 2007. A phylostratigraphy approach to uncover the genomic
633 history of major adaptations in metazoan lineages. *Trends Genet* 23:533-539.
634 Emera D, Yin J, Reilly SK, Gockley J, Noonan JP. 2016. Origin and evolution of developmental
635 enhancers in the mammalian neocortex. *Proc Natl Acad Sci U S A* 113:E2617-2626.
636 ENCODE. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57-
637 74.
638 Gaiti F, Jindrich K, Fernandez-Valverde SL, Roper KE, Degnan BM, Tanurdžić M. 2017. Landscape
639 of histone modifications in a sponge reveals the origin of animal cis-regulatory complexity. *eLife* 6.
640 Haberle V, Stark A. 2018. Eukaryotic core promoters and the functional basis of transcription
641 initiation. *Nat Rev Mol Cell Biol* 19:621-637.
642 He B, Chen C, Teng L, Tan K. 2014. Global view of enhancer-promoter interactome in human cells.
643 *Proc Natl Acad Sci U S A* 111:E2191-E2199.
644 Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C,
645 Ching KA, et al. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and
646 enhancers in the human genome. *Nat Genet* 39:311-318.
647 Ingolia NT. 2014. Ribosome profiling: new views of translation, from single codons to genome scale.
648 *Nat Rev Genet* 15:205-213.
649 Ingolia NT, Brar GA, Stern-Ginossar N, Harris MS, Talhouarne Gaëlle JS, Jackson SE, Wills MR,
650 Weissman JS. 2014. Ribosome Profiling Reveals Pervasive Translation Outside of Annotated Protein-
651 Coding Genes. *Cell Rep* 8:1365-1379.
652 Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. *Genome Res* 20:1313-
653 1326.
654 Kapranov P, Willingham AT, Gingeras TR. 2007. Genome-wide transcription and the implications for
655 genomic organization. *Nature Rev Genet* 8:413-423.
656 Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler aD. 2002. The
657 Human Genome Browser at UCSC. *Genome Res* 12:996-1006.

658 Kherdjemil Y, Lalonde RL, Sheth R, Dumouchel A, de Martino G, Pineault KM, Wellik DM, Stadler
659 HS, Akimenko M-A, Kmita M. 2016. Evolution of Hoxa11 regulation in vertebrates is linked to the
660 pentadactyl state. *Nature* 539:89-92.
661 Kim SK, Mayer MG, Rödelsperger C, Witte H, Riebesell M, Sommer RJ. 2015. The orphan gene
662 dauerless regulates Dauer development and intraspecific competition in nematodes by copy number
663 variation. *PLoS Genet* 11.
664 Kim T-K, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-
665 Haley K, Kuersten S, et al. 2010. Widespread transcription at neuronal activity-regulated enhancers.
666 *Nature* 465:182-187.
667 Kratochwil CF, Liang Y, Gerwin J, Woltering JM, Urban S, Henning F, Machado-Schiaffino G,
668 Hulsey CD, Meyer A. 2018. Agouti-related peptide 2 facilitates convergent evolution of stripe
669 patterns across cichlid fish radiations. *Science* 362:457-460.
670 Kryuchkova-Mostacci N, Robinson-Rechavi M. 2015. Tissue-specific evolution of protein coding
671 genes in human and mouse. *PLoS One* 10:e0131673.
672 Kvon EZ, Kamneva OK, Melo US, Barozzi I, Osterwalder M, Mannion BJ, Tissières V, Pickle CS,
673 Plajzer-Frick I, Lee EA, et al. 2016. Progressive loss of function in a limb enhancer during snake
674 evolution. *Cell* 167:633-642.e611.
675 Kvon EZ, Kazmar T, Stampfel G, Yáñez-Cuna JO, Pagani M, Schernhuber K, Dickson BJ, Stark A.
676 2014. Genome-scale functional characterization of Drosophila developmental enhancers in vivo.
677 *Nature* 512:91-95.
678 Levine M, Cattoglio C, Tjian R. 2014. Looping back to leap forward: transcription enters a new era.
679 *Cell* 157:13-25.
680 Li B, Qing T, Zhu J, Wen Z, Yu Y, Fukumura R, Zheng Y, Gondo Y, Shi L. 2017. A Comprehensive
681 Mouse Transcriptomic BodyMap across 17 Tissues by RNA-seq. *Sci Rep* 7.
682 Li D, Yan Z, Lu L, Jiang H, Wang W. 2014. Pleiotropy of the de novo-originated gene MDF1. *Sci*
683 *Rep* 4.
684 Li W, Notani D, Rosenfeld MG. 2016. Enhancers as non-coding RNA transcription units: recent
685 insights and future perspectives. *Nature Rev Genet* 17:207-223.
686 Lizio M, Harshbarger J, Shimoji H, Severin J, Kasukawa T, Sahin S, Abugessaisa I, Fukuda S, Hori
687 F, Ishikawa-Kato S, et al. 2015. Gateways to the FANTOM5 promoter level mammalian expression
688 atlas. *Genome Biol* 16.
689 Marletaz F, Firbas PN, Maeso I, Tena JJ, Bogdanovic O, Perry M, Wyatt CDR, de la Calle-Mustienes
690 E, Bertrand S, Burguera D, et al. 2018. Amphioxus functional genomics and the origins of vertebrate
691 gene regulation. *Nature* 564:64-70.
692 McLysaght A, Hurst LD. 2016. Open questions in the study of de novo genes: what, how and why.
693 *Nature Rev Genet* 17:567-578.

694 Medina-Rivera A, Santiago-Algarra D, Puthier D, Spicuglia S. 2018. Widespread Enhancer Activity
695 from Core Promoters. *Trends Biochem Sci* 43:452-468.
696 Michel AM, Fox G, M. Kiran A, De Bo C, O'Connor PBF, Heaphy SM, Mullan JPA, Donohue CA,
697 Higgins DG, Baranov PV. 2014. GWIPS-viz: development of a ribo-seq genome browser. *Nucleic*
698 *Acids Res* 42:D859-D864.
699 Neme R, Tautz D. 2016. Fast turnover of genome transcription across evolutionary time exposes
700 entire non-coding DNA to de novo gene emergence. *eLife* 5.
701 Neme R, Tautz D. 2013. Phylogenetic patterns of emergence of new genes support a model of
702 frequent de novo evolution. *BMC Genomics* 14.
703 Noguchi S, Arakawa T, Fukuda S, Furuno M, Hasegawa A, Hori F, Ishikawa-Kato S, Kaida K, Kaiho
704 A, Kanamori-Katayama M, et al. 2017. FANTOM5 CAGE profiles of human and mouse samples. *Sci*
705 *Data* 4.
706 Osterwalder M, Barozzi I, Tissières V, Fukuda-Yuzawa Y, Mannion BJ, Afzal SY, Lee EA, Zhu Y,
707 Plajzer-Frick I, Pickle CS, et al. 2018. Enhancer redundancy provides phenotypic robustness in
708 mammalian development. *Nature* 554:239-243.
709 Partha R, Chauhan BK, Ferreira Z, Robinson JD, Lathrop K, Nischal KK, Chikina M, Clark NL.
710 2017. Subterranean mammals show convergent regression in ocular genes and enhancers, along with
711 adaptation to tunneling. *eLife* 6.
712 Peter IS, Davidson EH. 2011. Evolution of gene regulatory networks controlling body plan
713 development. *Cell* 144:970-985.
714 Pliner HA, Packer JS, McFaline-Figueroa JL, Cusanovich DA, Daza RM, Aghamirzaie D, Srivatsan
715 S, Qiu X, Jackson D, Minkina A, et al. 2018. Cicero Predicts cis-Regulatory DNA Interactions from
716 Single-Cell Chromatin Accessibility Data. *Mol Cell* 71:858-871.e858.
717 Prabh N, Rödelsperger C. 2016. Are orphan genes protein-coding, prediction artifacts, or non-coding
718 RNAs? *BMC Bioinform* 17.
719 Prud'homme B, Gompel N, Carroll SB. 2007. Emerging principles of regulatory evolution. *Proc Natl*
720 *Acad Sci U S A* 104:8605-8612.
721 Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features.
722 *Bioinformatics* 26:841-842.
723 Roscito JG, Sameith K, Parra G, Langer BE, Petzold A, Moebius C, Bickle M, Rodrigues MT, Hiller
724 M. 2018. Phenotype loss is associated with widespread divergence of the gene regulatory landscape in
725 evolution. *Nature Commun* 9.
726 Ruiz-Orera J, Alba MM. 2019. Translation of small open reading frames: roles in regulation and
727 evolutionary innovation. *Trends Genet* 35:186-198.
728 Ruiz-Orera J, Messeguer X, Subirana JA, Alba MM. 2014. Long non-coding RNAs as a source of
729 new peptides. *eLife 3*.

730 Ruiz-Orera J, Verdaguer-Grau P, Villanueva-Canas JL, Messeguer X, Alba MM. 2018. Translation of
731 neutrally evolving peptides provides a basis for de novo gene evolution. *Nat Ecol Evol* 2:890-896.
732 Schmitz JF, Ullrich KK, Bornberg-Bauer E. 2018. Incipient de novo genes can evolve from frozen
733 accidents that escaped rapid transcript turnover. *Nat Ecol Evol* 2:1626-1632.
734 Schwaiger M, Schonauer A, Rendeiro AF, Pribitzer C, Schauer A, Gilles AF, Schinko JB, Renfer E,
735 Fredman D, Technau U. 2014. Evolutionary conservation of the eumetazoan gene regulatory
736 landscape. *Genome Res* 24:639-650.
737 Sebé-Pedrós A, Chomsky E, Pang K, Lara-Astiaso D, Gaiti F, Mukamel Z, Amit I, Hejnol A, Degnan
738 BM, Tanay A. 2018. Early metazoan cell type diversity and the evolution of multicellular gene
739 regulation. *Nat Ecol Evol* 2:1176-1188.
740 Sebé-Pedrós A, Saudemont B, Chomsky E, Plessier F, Mailhe MP, Renno J, Loe-Mie Y, Lifshitz A,
741 Mukamel Z, Schmutz S, et al. 2018. Cnidarian cell type diversity and regulation revealed by whole-
742 organism single-cell RNA-Seq. *Cell* 173:1520-1534 e1520.
743 Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A,
744 Nakamura M, Arakawa T, et al. 2003. Cap analysis gene expression for high-throughput analysis of
745 transcriptional starting point and identification of promoter usage. *Proc Natl Acad Sci U S A*
746 100:15776-15781.
747 Spitz F, Furlong EEM. 2012. Transcription factors: from enhancer binding to developmental control.
748 *Nat Rev Genet* 13:613-626.
749 Tabula Muris Consortium. 2018. Single-cell transcriptomics of 20 mouse organs creates a Tabula
750 Muris. *Nature* 562:367-372.
751 Tautz D, Domazet-Loso T. 2011. The evolutionary origin of orphan genes. *Nat Rev Genet* 12:692-
752 702.
753 Teichmann SA, Babu MM. 2004. Gene regulatory network growth by duplication. *Nature Genet*
754 36:492-496.
755 Van Oss SB, Carvunis A-R. 2019. De novo gene birth. *PLoS Genet* 15.
756 Villar D, Berthelot C, Aldridge S, Rayner Tim F, Lukk M, Pignatelli M, Park Thomas J, Deaville R,
757 Erichsen Jonathan T, Jasinska Anna J, et al. 2015. Enhancer evolution across 20 mammalian species.
758 *Cell* 160:554-566.
759 Warnefors M, Eyre-Walker A. 2011. The accumulation of gene regulation through time. *Genome Biol*
760 *Evol* 3:667-673.
761 Werner MS, Sieriebriennikov B, Prabh N, Loschko T, Lanz C, Sommer RJ. 2018. Young genes have
762 distinct gene structure, epigenetic profiles, and transcriptional regulation. *Genome Res* 28:1675-1687.
763 Willemsen A, Félez-Sánchez M, Bravo IG, Bapteste E. 2019. Genome plasticity in papillomaviruses
764 and de novo emergence of E5 oncogenes. *Genome Biol Evol* 11:1602-1617.
765 Wilson BA, Masel J. 2011. Putatively Noncoding Transcripts Show Extensive Association with
766 Ribosomes. *Genome Biol Evol* 3:1245-1252.

767 Witt E, Benjamin S, Svetec N, Zhao L. 2019. Testis single-cell RNA-seq reveals the dynamics of de
768 novo gene transcription and germline mutational bias in Drosophila. *eLife* 8.
769 Wu X, Sharp PA. 2013. Divergent transcription: a driving force for new gene origination? *Cell*
770 155:990-996.
771 Zhang J, Chen J-Y, Shen QS, Zhou W-Z, Peng J, He BZ, Li Y, Liu C-J, Luan X, Ding W, et al. 2015.
772 Emergence, Retention and Selection: A Trilogy of Origination for Functional De Novo Proteins from
773 Ancestral LncRNAs in Primates. *PLoS Genet* 11.
774 Zhang L, Ren Y, Yang T, Li G, Chen J, Gschwend AR, Yu Y, Hou G, Zi J, Zhou R, et al. 2019.
775 Rapid evolution of protein diversity by de novo origination in *Oryza*. *Nat Ecol Evol* 3:679-690.
776

Supplementary figures

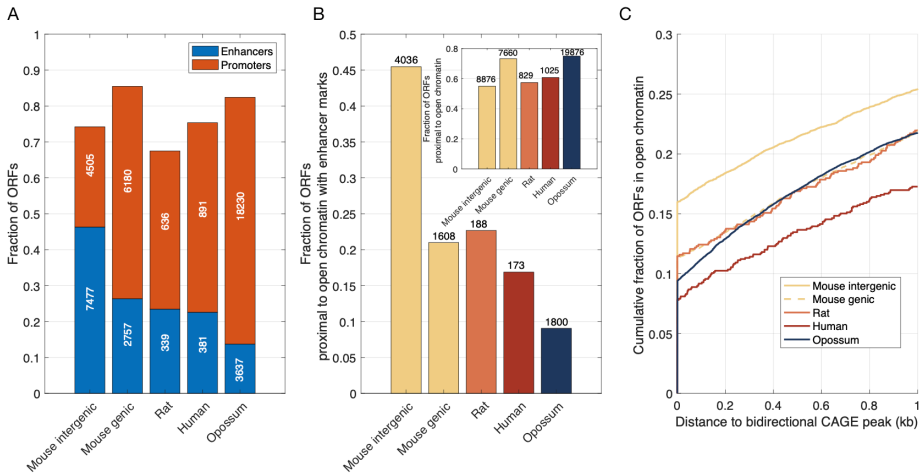


Figure S1. Exon length is not responsible for the proximity of young ORFs to enhancers. Repeating the analyses shown in Fig. 1B-D using windows of 1.5 kb around the central nucleotide of the first exon of each ORF, rather than the first nucleotide of the first exon, results in the same qualitative trend that mouse-specific intergenic ORFs are more likely to be proximal to enhancers than the other classes of ORFs studied here.

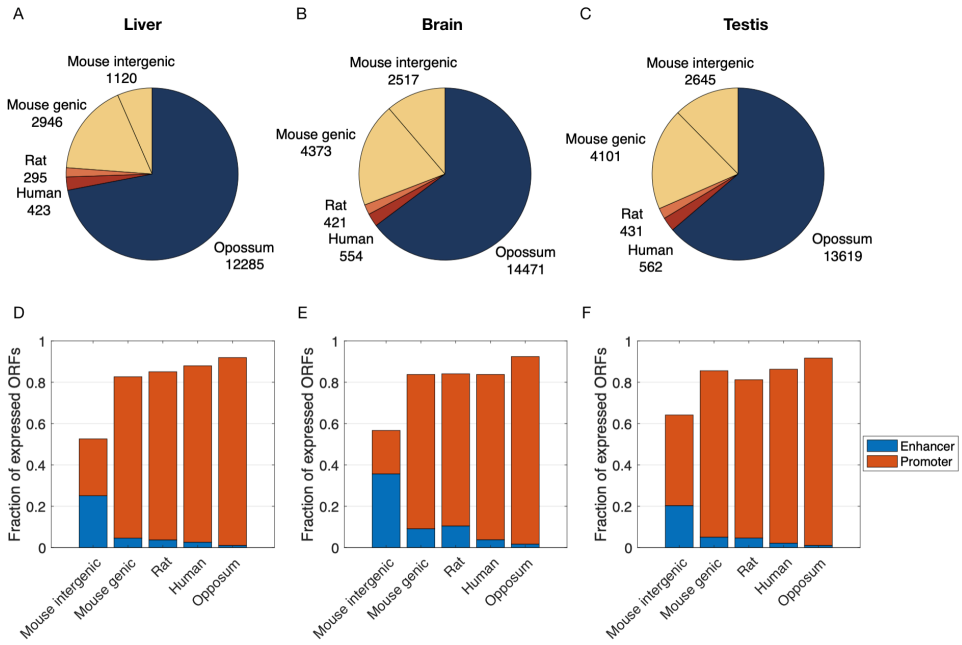


Figure S2. Mouse-specific intergenic ORFs are more associated with active enhancers in individual tissues than older and genic ORFs. A-C) Number of ORFs of each class that are in regions of open chromatin and that are expressed (FPKM > 0) in (A) liver, (B) brain, and (C) testis. D-F) Fraction of ORFs from each class (as presented in A-C) that are proximal to promoters or enhancers in (D) liver, (E) brain, and (F) testis.

Figure S3. Enhancer redundancy may help prevent conflicts in the repurposing of enhancers to promoters. Visualization of 6 genomic regions that contain 7 mouse-specific intergenic ORFs that are proximal to promoters and show evidence of being repurposed enhancers.

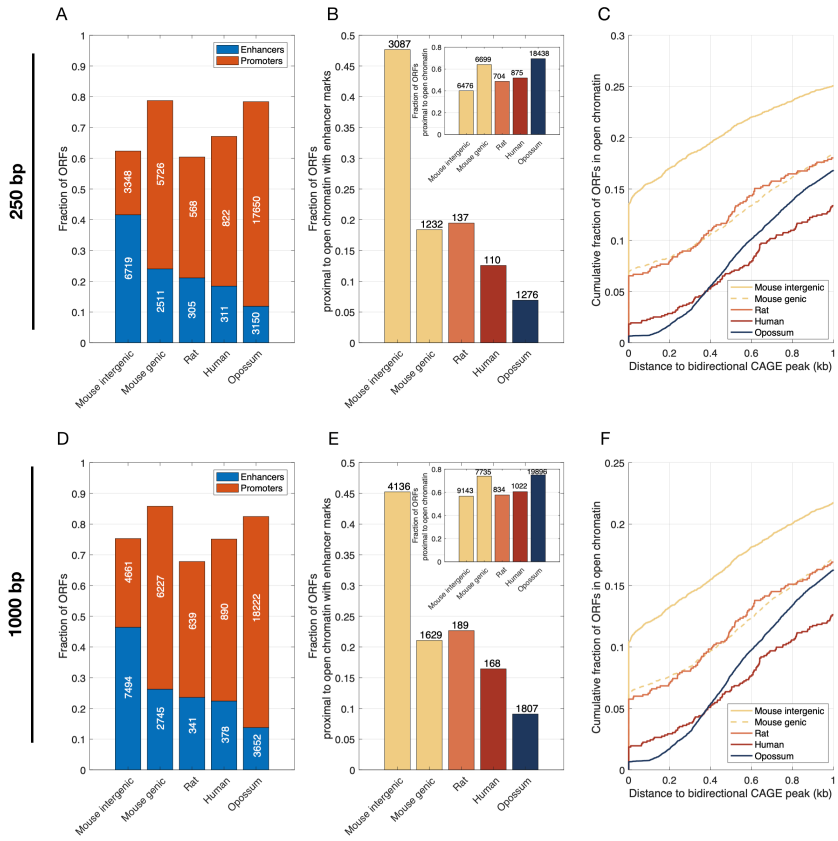


Figure S4. Changing our definition of proximal from within 500bp to within (A-C) 250bp or (D-F) 1000bp does not qualitatively alter the trends shown in Fig. 1B-D.

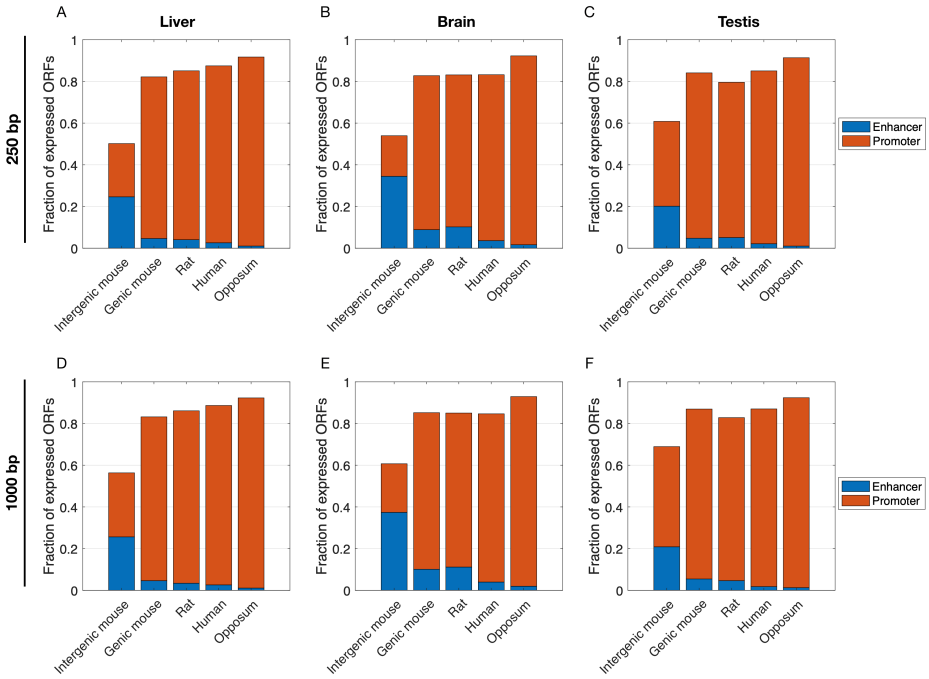


Figure S5. Changing our definition of proximal from within 500bp to within (A-C) 250bp or (D-F) 1000bp does not qualitatively alter the trends shown in Fig. S2D-F.

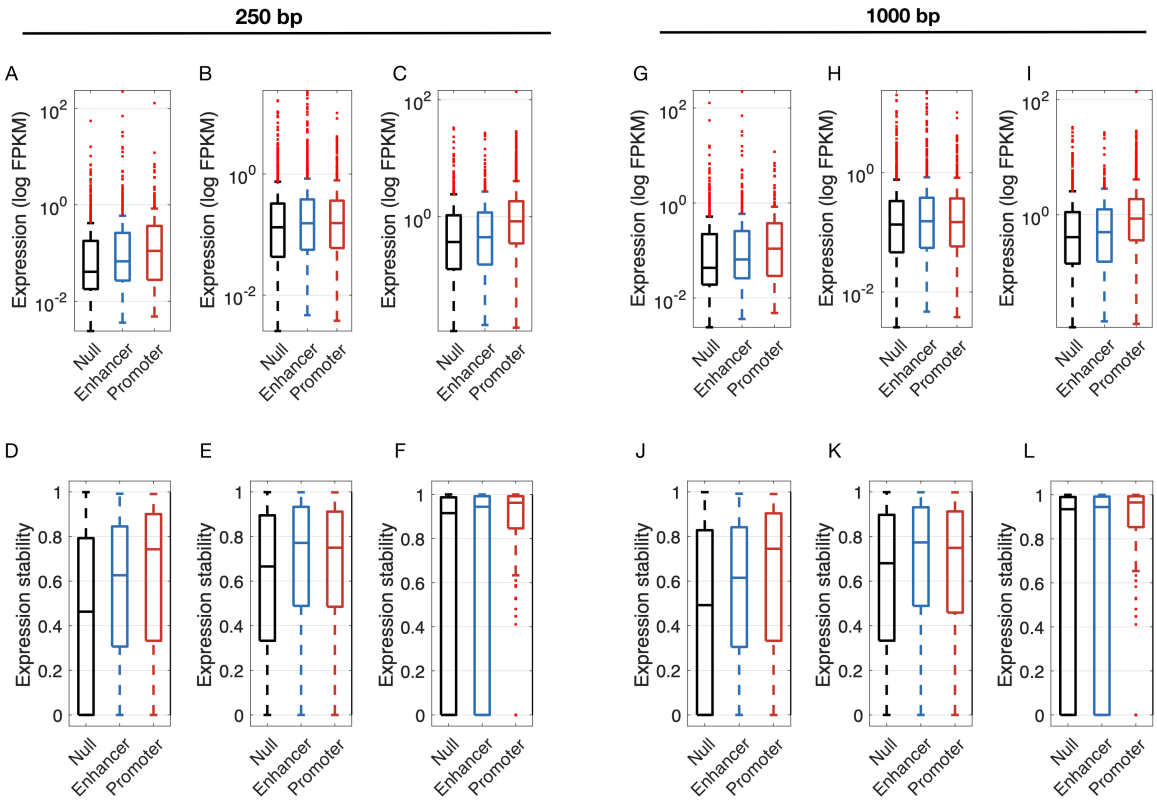


Figure S6. Changing our definition of proximal from within 500bp to within (A-F) 250bp or (G-L) 1000bp does not qualitatively alter the trends shown in Fig. 2.

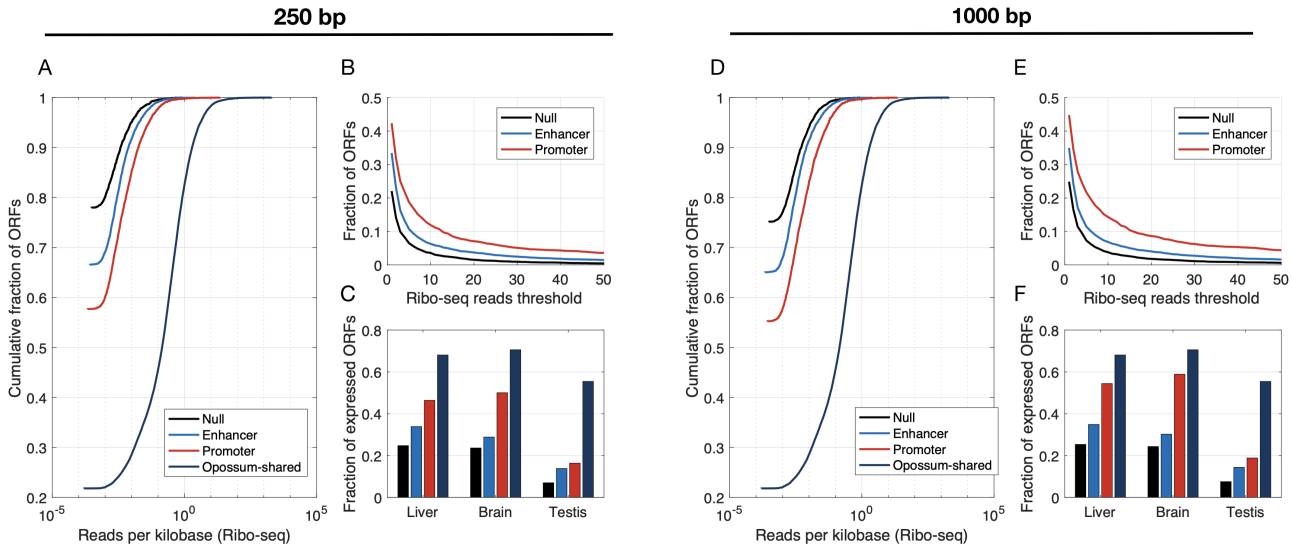


Figure S7. Changing our definition of proximal from within 500bp to within (A-C) 250bp or (D-F) 1000bp does not qualitatively alter the trends shown in Fig. 3.

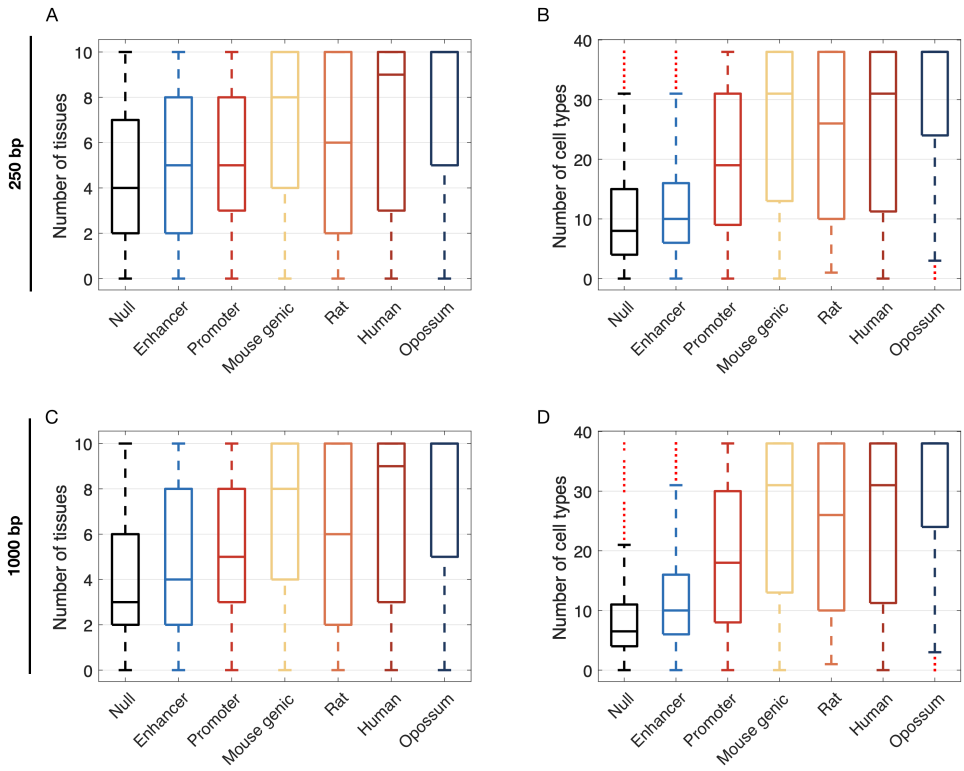


Figure S8. Changing our definition of proximal from within 500bp to within (A,B) 250bp or (C,D) 1000bp does not qualitatively alter the trends shown in Fig. 4.