False-Positive IRESes from *Hoxa9* and other genes resulting from errors in mammalian 5' UTR annotations

Christina Akirtava[a,1], Gemma E. May[a,1], C. Joel McManus[a,b,*].

[a]Department of Biological Sciences, [b]Computational Biology Department, Carnegie Mellon University, Pittsburgh PA 15213

[1] These authors contributed equally to this study
* Correspondence to C. Joel McManus, mcmanus@andrew.cmu.edu

Christina Akirtava 0000-0001-7587-2303
Charles Joel McManus 0000-0002-6605-2642
Gemma Ean May 0000-0003-1583-2400

Author Contributions: C.A., G.E.M., and C. J. M. designed research. C.A. and G.E.M. performed research and analyzed data. C.J.M. wrote the manuscript, with input from all authors.

Keywords: Hox, IRES, UTR, Translation, Bicistronic

Abstract

Hyperconserved genomic sequences have great promise for understanding core biological processes. It has been recently proposed that scores of hyperconserved transcript leaders (hTLs) encode Internal Ribosome Entry Sites (IRESes) that drive cap-independent translation in part via interactions with ribosome expansion segments. However, the direct functional significance of such interactions has not yet been definitively demonstrated. We provide evidence that the putative IRESes previously reported in *Hoxa* gene hTLs are not transcribed. Instead, these regions function independently as transcriptional promoters. In addition, we find the proposed RNA structure of the putative *Hoxa9* IRES is not conserved. Instead, sequences previously shown to be essential for putative IRES activity encode a hyperconserved transcription factor binding site (E-box) that contributes to its promoter activity by binding to the transcription factors *USF1* and *USF2*. Similar E-box sequences enhance the promoter activities of other putative *Hoxa* gene IRESes. Moreover, we provide evidence that the vast majority of hTLs with putative IRES activity overlap transcriptional promoters, enhancers, and 3' splice sites that are most likely responsible for their reported IRES activities. These results argue strongly against recently reported widespread IRES-like activities from hTLs and contradict proposed interactions between ribosomal expansion segment ES9S and putative IRESes. Our work underscores the importance of accurate transcript annotations, controls in bicistronic reporter assays, and the power of synthesizing publicly available data from multiple sources.

Introduction

As a critical step in gene expression, the translation of mRNA into protein is highly regulated. Eukaryotic translation is primarily controlled at the initiation stage, in which ribosomes identify start codons and begin synthesizing protein (1, 2). During proliferative growth, most mRNA translation initiates through a cap-dependent mechanism in which the 5' 7-mG interacts with initiation factors to recruit a pre-initiation complex (PIC) composed of the 40S small ribosomal subunit and multiple initiation factors. Once recruited, PICs scan directionally 5' to 3' until a start codon is recognized, the large ribosomal subunit is recruited, and translation commences. Under stress conditions, this cap-dependent translation is largely repressed due to inactivation of initiation factors. In such circumstances, ribosomes can be recruited to mRNA through cap-independent mechanisms, including Internal Ribosome Entry Sites (IRESes). IRESes are often found in viruses, as these pathogens often suppress cap-dependent translation of cellular RNAs to commandeer ribosomes for viral protein synthesis. IRESes have also been reported in cellular mRNA, though their roles in translation remain controversial (1, 2).

Several studies have coalesced on a surprising model in which hyperconserved transcript leaders include IRES-like sequences that drive cap-independent translation in specific cell types during development. These IRES-like elements were first proposed for mammalian *Hoxa* genes, based on the observation that the annotated mouse transcript leaders from several *Hoxa* genes drove expression in bicistronic luciferase assays, a classic test for cap-independent translation (3). It has also been proposed that ribosome expansion segment 9S (ES9S), a stem loop that protrudes from the ribosome, binds to a structured stem loop in the *Hoxa9* IRES-like sequence to recruit ribosomes to the *Hoxa9* transcript (4). Mammalian ES9S was also shown to bind G-rich motifs found in many mRNAs *in vitro*, which was proposed to drive cap-independent translation of many cellular transcripts (5).

The possibility of widespread cap-independent translation driven by interactions between IRESes and expansion segments is tantalizing. However, previously reported IRESes in cellular mRNA have been fraught with controversy (6–8). In particular, the bicistronic luciferase reporter system used to support *Hoxa* IRES activity is widely known to produce false positive results due to transcriptional promoter activity, cryptic splicing, and ribosomal readthrough of the upstream cistron (9–14)(Figure S1). Consequently, the exact boundaries of transcript leaders must be used to prevent the addition of

upstream promoters and 3' splice sites. Indeed, it has been noted that the Hox genes likely have much shorter transcript leaders than those used in bicistronic IRES assays (7). Furthermore, previous RNAi control experiments suggested that putative *Hoxa* gene IRESes have independent promoter activity. While siRNA targeting the upstream Rluc eliminated Rluc expression,~ 30% of Fluc expression resisted RNAi, indicating substantial monocistronic Fluc transcripts (3). However, the authors of the study drew the opposite conclusion.

The proposed functional interaction between the *Hoxa9* P4 stemloop and ribosome ES9S is also problematic. Previous work found sequences complementary to human ES9S did not support IRES activity (15). In addition, the proposed IRES RNA structures are inconsistent with functional assays. A cryo-EM structure model of this interaction appears to show the helices oriented as kissing stem loops. This structure most likely involves base pairing between nucleotides in the G-rich P4 loop with the C-rich loop of ES9S (Figure S2). However, mutations to the G-rich loop of P4 did not disrupt its apparent IRES activity (4, 5), suggesting that this proposed interaction is dispensable. Instead, the authors depicted a side-by-side helix interaction model inconsistent with their cryo-EM structure. Although some mutations that disrupt the P4 stem greatly reduced apparent IRES activity, compensatory mutations to restore P4 base pairing did not restore IRES function (4). Finally, a deletion of the 5' half of the P4 stem loop did not disrupt IRES-like activity in the bicistronic reporter. Together, these observations cast doubt on the model that ES9S binds to the *Hoxa9* P4 stem loop to drive cap-independent translation.

Recently, a high-throughput analysis defined a set of 589 hyperconserved transcript leaders (hTLs) with strong enrichment for genes involved in mammalian development (16). Hundreds of these hTLs were tested for IRES-like activities in bicistronic reporter assays and thirty-seven percent (90 / 241) drove substantial expression of the downstream luciferase cistron, suggesting that hTLs may frequently encode IRES-like functional elements. However, the possibility that these putative IRES activities may instead reflect functional promoter elements or cryptic 3' splices sites was not directly investigated. While the authors showed *Fluc* / *Rluc* protein and RNA ratios were not strongly correlated, this could result from variance in luciferase and qPCR measurements. Consequently, though they provide an alluring model for new modes of translational control during mammalian development, the authenticity of hTL IRES-like elements has not been conclusively established.

In this work, we investigated the possibility that putative IRES-like elements in mammalian hTLs actually encode transcriptional promoters and 3' splice sites. We show that the putative *Hoxa9* IRES has no signs of structural conservation and does not appear to be expressed in mature mRNA. Instead, sequences encoding putative IRESes from mouse *Hoxa* genes act as independent promoters. In addition, we demonstrate that a sequence previously identified as essential for *Hoxa9* IRES activity is a classical "E-box" site recognized by the bHLH transcription factors *USF1*, *USF2*, *MYC*, and *MAX*. Putative IRESes from other *Hoxa* genes similarly have conserved E-box motifs that contribute to their promoter activities. Furthermore, the proposed IRES-like elements in the transcript leaders of *Chrdl1*, *Cnot3*, *Cryab* and *Slc25a14* also have strong promoter activities. We also find putative hTLs frequently overlap other functional elements, including protein coding sequences, which could explain their conservation. Finally, we show that recently proposed IRES-like hTLs are overwhelmingly further enriched in annotated promoters, 3' splice sites, and internal transcription initiation, and these elements can be used to accurately predict their reported IRES-like activities.

Results

We first investigated the putative IRES region of *Hoxa9*, which has been called the paradigmatic example hyperconserved transcript leader (16). Many IRESes, including viral IRESes, fold into complex functional RNA structures. Previous studies reported a complex secondary structure for the *Hoxa9* IRES based on SHAPE probing (3). Two RNA base pairing regions (P3a and P4) were required for IRES activity (3, 4). The P4 stem-loop was later shown to interact with ribosomal expansion segment ES9S *in vitro* (4), yet the P4 structure was not required for bicistronic reporter activity (4). Thus, the functional significance of the P4 region remains unclear. To further investigate the importance of the *Hoxa9* IRES structure, we examined its conservation. The putative IRES regions of both mouse and zebrafish *Hoxa9* were previously shown to drive bicistronic reporter expression in mouse tissue culture (3). However, the predicted RNA structures of mouse, human and zebrafish *Hoxa9* IRES regions differ substantially, such that the P3a domain is not predicted to form in the homologous human sequence, and neither P3a or P4 are likely to form in zebrafish (Figure 1A & B). Furthermore, we found no evidence of significant structural covariation in RNA sequence alignments from 230 vertebrates (17) using Infernal (18) and R-scape

(19)(Figure 1C), despite having enough statistical power to detect such pairing (20) (Table S1). These results indicate that the proposed RNA structure of the *Hoxa9* IRES region, including domains previously reported to be essential for IRES activity, is not evolutionarily constrained.

To drive cap-independent translation, the mouse *Hoxa9* IRES must be included in its TL. Previous work noted that this may not be the case (7). Our evaluation of public RNA-seq data indicates that the annotated TL of mouse *Hoxa9* used in previous studies shows little evidence of transcription in mouse tissues (Figure 2A). In nearly all tissues analyzed, including embryonic neural tube, the tissue from which the IRES was first reported (3), ENCODE RNA-seq data show negligible levels of transcribed RNA in the upstream region of the annotated 5' UTR. Instead, transcript levels sharply increase close to the *Hoxa9* start codon, immediately downstream of a strong transcription start site annotated in the refTSS database. These short 5' UTR isoforms are also supported by ENCODE long-read RNA-seq data from developing embryos (Figure 2A). Although the extended TL is partially supported by one long read, this appears to be an unspliced intron from an upstream fusion transcript (Figure 2A). A very similar set of isoforms is supported by human RNA-seq data (Figure S3). Thus, the extended *Hoxa9* TL, including the region sufficient for IRES activity (3), is generally not transcribed in mice or humans.

As the putative *Hoxa9* IRES did not appear to be transcribed, we examined an alternative hypothesis that the reported UTR region encodes functional DNA elements. Manual examination revealed the extended TL region overlapped two enhancers and a promoter annotated by the ENCODE consortium and the Eukaryotic Promoter Database (21, 22)(Figure 2A). We tested this region for promoter activity in mouse C3H/10T1/2 embryonic mesenchymal cells, which were previously used to study the Hoxa9 IRES (3). To most directly compare our results to previous IRES studies, we used a modified bicistronic reporter plasmid lacking the upstream SV40 promoter (pRF-ΔSV40) and cloned putative *Hoxa9* promoter regions between Rluc and Fluc (Figure 2B). Strikingly, both the putative extended TL and putative IRES regions of mouse and human *Hoxa9* drove expression of Firefly luciferase. Additionally, this expression was absent in reporters in which the putative IRES region was reversed (Figure 2B). These results show that the putative IRES region of *Hoxa9* encodes a functional promoter.

Based on bicistronic reporter assays, IRES-like activities were previously reported for the annotated TLs from other Hox genes, including *Hoxa3, a4, a5, a7,* and *a11*, while *Hoxa1*, *a2*, *a6*, *a10*,

and *a13* did not show activity (3, 16). We next investigated whether the IRES-like regions reported for these genes also have independent promoter activity. Remarkably, all of the previously reported IRES-like TLs we tested (*Hoxa3*, *a5*, *a7*, and *a11*) drove expression of *Fluc* independent of an upstream SV40 promoter, while the non-IRES TLs (*Hoxa1*, *a2*, *a6*, *a10*, and *a13*) had lower *Fluc*/*Rluc* ratios indistinguishable from background noise (Figure 2C; see also Figure S4). Together, these results suggest the previously reported *Fluc* expression in bicistronic reporter plasmids containing upstream sequences from *Hoxa* genes were due to monocistronic *Fluc* transcripts driven by independent promoters, and not from bonafide IRESes.

We next considered previously identified critical sequences in the *Hoxa9* P4 region. Sequences in the highly conserved 3' half of this region were previously shown to be required for bicistronic reporter activity and normal skeletal development (3, 4). Specifically, these were sensitive to mutations in the nucleotides underlined "GA**CACGUGA**C", and similar sequences can be found in other putative *Hoxa* gene IRESes (4). Using FIMO (23) to search for transcription factor binding sites, we found the equivalent DNA sequence (CACGTG) matches the E-box motif recognized by bHLH transcription factors *MYC*/*MAX*(24) and *USF1*/*USF2* (25, 26). Recent work showed *USF2* binds upstream of human *Hoxa9*, and that co-depletion of *USF1* and *USF2* decreases *Hoxa9* expression in human tissue culture cells (27)(Figure 3A). Remarkably, this E-box appears to be universally conserved in vertebrates, with the only substitution being CACATG, which also functions as an E-box. Another common TF binding site, the CAAT box, is universally conserved adjacent to the E-box. Notably, the G-rich sequence found in the loop region of the mouse *Hoxa9* P4 domain, which has a strong propensity to pair with mouse ES9S through a kissing stem-loop interaction (Figure S1), is not similarly conserved in vertebrates. Thus, sequences reported to be essential for mouse *Hoxa9* IRES-like activity encode a deeply conserved transcription factor binding site recently shown to drive expression of human *Hoxa9*.

We tested the hypothesis that these E-box motifs contribute to the promoter activity of *Hox* gene putative IRES regions. The promoter elements cloned between Fluc and Rluc in pRF-ΔSV40 plasmids appear to enhance spurious transcription of the upstream *Rluc* gene to various extents (Figure S4), which could complicate interpretations of the effect of mutations on *Fluc* expression. Thus, we placed *Rluc* and *Fluc* on separate plasmids to assay the importance of E-box sites in mammalian *Hoxa* genes (Fig 3B).

Mutating the E-box motifs reduced the promoter activity of mouse *Hoxa3, a5, a7, a11,* and mouse and human *Hoxa9* IRES regions (Figure 3C-D). Furthermore, siRNA co-depletion of mouse *USF1* and *USF2* led to a significant reduction in luciferase expression from wildtype mouse and human *Hoxa9* reporters, but not from reporters in which the E-box had been mutated (Figure 3C; $P<0.05$; 1-tail Welch's t-test). In contrast, *USF1*/*2* co-depletion did not reduce expression driven by other mouse *Hoxa* putative IRESes (Figure 2C), suggesting their E-boxes may be primarily regulated by *MYC/MAX* instead. We conclude that the IRES-like regions of mouse *Hoxa* genes encode functional E-boxes. The function of these sequences as E-boxes explains their necessity for bicistronic reporter expression in previous studies of putative *Hoxa9* IRES activity.

Using the *Hoxa9* gene as a prototypical example of a hyperconserved transcript leader, a recent study identified 589 hTLs in the mouse genome. The authors tested over two hundred of these elements in the bicistronic reporter system and reported ninety (37%) had IRES-like activity (16). Given the misannotation of *Hoxa* gene TLs, we next considered the possibility that these IRES-like hTLs may also be misannotated and encode functional promoters or 3' splice sites, which both give false-positive results in bicistronic reporter assays. Evaluation of annotated promoter elements (21, 22), transcription start sites (28), annotated splice sites, and short- and long-read RNA-seq data (29) for TLs reported to have such IRES-like activities revealed the vast majority (85 of 90; 94%) have promoter and / or splicing elements. For example, the *Dedd* gene TL, reported to have the highest IRES-like activity, overlaps two ENCODE promoters and two refTSS transcription start sites. ENCODE short- and long-read RNA-seq data support transcription initiation inside the transcript leader, with almost no evidence of full-length hTL expression (Figure 4A). Similarly, the second most active TL, *Ptp4a1*, appears to be completely misannotated, such that transcription always initiates at an EPD annotated promoter encoded within the hTL. The IRES-like TL of *Chrdl1* also appears to be misannotated and encodes two EPD promoters with refTSS sites supporting internal transcription initiation (Figure 4A). Other hTLs with reported IRES-like activity overlap 3' splice sites (Figure S5A and C). Finally, we tested hTLs from four mouse genes that were reported to have IRES activity in C3H/10T1/2 cells (16) - *Chrdl1*, *Cryab*, *Cnot3*, and *Slc25a14*. All showed independent promoter activity in pRF-ΔSV40 (Figure 4B). These results suggest that recently reported hTLs with IRES-like activity may also be misannotated, such that their activities in bicistronic reporter

assays result from monocistronic transcripts generated from internal promoters and 3' splice sites rather than bonafide cap-independent translation.

We next systematically evaluated the 589 previously reported hTLs for potential misannotation due to overlapping promoters, enhancers, transcription start sites, 3' splice sites, and protein coding sequences (CDSs), all of which could contribute to high conservation rates unrelated to translational control. Using combined promoter/enhancer sets, we find 93% of hTLs overlap ENCODE (509) and/or EPD (463) promoters (Figure S5B). Of those, 221 hTLs contain full-length EPD and 100 contain full-length ENCODE promoters. To further evaluate the accuracy of mouse TL annotations, we examined refTSS annotations and their underlying quantitative 5'-CAGE high-throughput sequencing data (28). Of the hTLs with public 5' CAGE data, only 40% have an annotated refTSS site within 10 nucleotides of their annotated 5' ends. Furthermore, 78% of all TSS-containing hTLs had the strongest CAGE peak within the hTL rather than near the 5' end. The complexity of the mouse transcriptome further complicates conclusions about TL conservation, as 17% of hTLs overlap annotated 3' splice sites and 43% of hTLs overlap annotated CDSs from alternative transcript isoforms (Figure S5C). Indeed, one-third of hTLs had CDS overlap covering at least 25% of their length and one-fifth were at least 50% overlapped by annotated CDSs (Figure S5D). These data show that many recently reported hTLs are misannotated such that their high conservation rates may reflect evolutionary pressure to maintain promoters, enhancers, splice sites, and protein coding sequences.

Our luciferase reporter assays show clear evidence that previously reported IRES-like elements result from transcriptional promoter activity. To further evaluate the potential for such false-positives in recently reported hTLs, we compared the frequencies of promoters and splice sites, and the distribution of 5' CAGE reads in hTLs that were reported to be "active" (90) and "non-active" (133) in the bicistronic assay. Active hTLs are 1.7-fold more likely to overlap at least one complete EPD promoter than "non-active" hTLs (53% vs 29%; Chi-squared test P = 0.0002107, Figure 4C). Similarly, annotated 3' splice sites were 2.8-fold enriched in active, as compared to non-active, hTLs (28% vs 10% Chi-squared test P = 0.000451, Figure 4C). Furthermore, ~83% of active TLs have their strongest CAGE peak within the TL compared to 68% for the non-active TLs (Chi-squared test P = 0.02009, Figure 4C). Finally, active hTLs have drastically more CAGE reads at internal refTSS sites (P = 0.00144), a lower fraction of CAGE reads

near their annotated 5' ends (P = 0.00173), and higher GC content (P = 2 x 10^{-6}) which is characteristic of promoter elements (Wilcoxon tests; Figure 4D). The enrichment of promoters, splice sites, and internal 5' CAGE reads in bicistronic "active" hTLs suggests that these elements generally drive bicistronic reporter expression through the creation of monocistronic *F. luc.* transcripts, rather than by cap-independent IRES-like activities.

If internal promoters and splice sites are responsible for the reported IRES-like activities in hTLs, we reasoned that such features could be used to predict their activities. To test this, we used logistic regression with these features to model their activity in the bicistronic reporter assay (Figure 4E-F; see Methods). Strikingly, this approach generated models that were up to 84% accurate at predicting IRES-like activity (mean accuracy 68%). Features associated with promoters (GC content, EPD promoter counts, E-boxes) and 3' splice sites were positively correlated with bicistronic active hTLs, while those reflecting high levels of full length hTL transcription (5' refTSS fraction and 5' RNA-seq bias) were associated with inactive hTLs. The ability to accurately predict bicistronic assay activity from these genomic features strongly supports the conclusion that such activities are false-positives, inconsistent with their putative functions as IRES-like elements driving cap-independent translation.

Discussion

The vast majority of mRNAs are believed to undergo cap-dependent translation in rapidly dividing cells, while cap-independent mechanisms, including IRESes, are used primarily during cell stress (2). Over the last decade, multiple studies have coalesced on an intriguing model proposing key developmental genes are regulated by cap-independent translation driven by IRES-like sequences and structures in hyperconserved 5' transcript leaders (hTLs). However, this model was founded on bicistronic reporter assays, which are subject to common false-positive results due to cryptic promoter and splicing activities. Here, we examined previously reported hTLs with putative IRES-like activities in *Hoxa9* and other genes. We found these TLs are often misannotated, such that their putative IRES regions are not transcribed. Instead, many of these false-positive IRESes encode transcriptional promoters. Consistent with this, a concurrent independent study found much shorter 5' UTRs in Hox genes expressed in developing mouse somites and showed the putative *Hoxa9* IRES has promoter activity (30). Furthermore, we successfully predicted putative IRES-like activity using known annotated promoters, 3' splice sites,

CAGE-seq 5' end data and public RNA-seq data. Finally, we find promoters, splice sites, enhancers, and even protein coding sequences overlap hTLs, which may explain their sequence conservation. Our results provide conventional explanations for unconventional results from previous studies, requiring a reevaluation of the proposal that these TLs drive cap-independent translation.

We found the putative *Hoxa9* IRES is not expressed in mouse RNA-seq data and instead encodes a functional promoter. Consistent with this, our R-scape analysis, made possible by recently published mammalian genome sequences (17), indicates the proposed RNA structure of the *Hoxa9* IRES (2015) is not constrained by evolution. Instead, the P4 region encodes a hyperconserved E-box motif recognized by USF1/2 whose mutation drastically decreases promoter activity. The recently reported M5 mutation (4), which eliminates this E-box ("CACGTGA" to "CAC**TATT**") decreased expression in the bicistronic reporter assay, as would be expected for loss of a transcription factor binding site. However, the M5 mutation also shifted *Hoxa9* mRNA from the polysome towards the monosome (4). While this may seem to support IRES-like elements, it can also be explained by promoter activity. Mutating the E-box likely decreases the production of the natural 82-nt TL isoform, such that spurious longer transcripts and unspliced *Hoxa9*/*a10* fusion transcripts make up a larger fraction of *Hoxa9* mRNA in polysome gradients. These longer transcripts include up to 14 uORFs and would not be translatable. Similar reasoning could explain how deletion of the putative IRES region alters *Hoxa9* polysome association in mouse embryos (3). Thus, the intrinsic promoter activity we observed in *Hoxa9* genes provides a conventional explanation for the effects of these mutations, especially since the putative IRES is not transcribed *in vivo* (Figs 1, S3, 30).

A recent study also proposed the *Hoxa9* P4 stem loop recruits translation preinitiation complexes (PICs) through interactions with ribosomal expansion segment ES9S (4). A transcriptome-wide screen reported that other G-rich mouse RNAs can similarly bind to ES9S in affinity purification assays (5). However, these assays were performed at 4°C and may not be physiologically relevant. Indeed, we show the P4 domain and ES9S have the potential to form kissing loops with nine G-C and one G-U base pairs (Figure S2), consistent with a published cryo-EM structure (4). This interaction has a predicted free energy of -15.94 kcal/mol (RNAcofold)(31) and would thus be very stable under cryo-EM and affinity purification conditions. Notably, neither the P4 structure nor the G-rich stretch were required for IRES-like

activity (4), and neither are evolutionarily conserved (Figures 1 and 3). Because the putative IRES structure is not conserved, and is rarely, if ever, expressed as a transcript leader (Figs 1, S3, 30), our results contradict the notion that mammalian ES9S recruits PICs to *Hoxa9* and other mRNAs for cap-independent translation. Instead, we propose that the interactions observed between the ES9S and G-rich mRNA are coincidental associations stabilized by low temperature.

The TLs of *Hoxa3*, *a4*, *a5*, *a7*, and *a11* were also previously reported to have IRES activity based on bicistronic assays (3, 16). As with *Hoxa9*, many of the other previously reported Hoxa gene IRES-like TLs appear to be misannotated. For example, Xue et al defined a 1,106 nt TL for *Hoxa4* using 5' RACE. However, the contemporaneous transcript annotation indicated a 15 nt leader, which is supported by RNA-seq data (Figure S6). Similarly, the 1,168 nt and 496 nt IRES-like TLs from mouse *Hoxa7* and *a11* appear to be only ~112 nt and ~90 nt long, respectively (Figure S6). Consistent with their misannotation, all the IRES-like *Hoxa* TLs we tested (*a3*, *a5*, *a7*, and *a11*) showed independent promoter activities, while non-IRES Hoxa TLs did not. Strikingly, sequences previously shown to be sufficient for putative *Hoxa3, a4, a5,* and *a11* IRES activities overlap annotated promoters and functional transcription start sites (Figure S6). Moreover, we found conserved E-boxes in all the *Hoxa* TLs with putative IRESes (Figure S6), but not in non-IRES TLs. Mutating these E-box sites decreased the strength of *Hoxa* IRES promoter activities, suggesting these elements are important regulators of many *Hoxa* genes. As with endogenous human *Hoxa9*, depletion of *USF1* and *USF2* decreased expression from the mouse and human *Hoxa9* reporters. The E-boxes from other *Hoxa* genes may be primarily recognized by *MYC/MAX*, as this heterodimer binds the same sequence (24). We conclude E-boxes are important motifs regulating *Hoxa* genes. Overall, our results suggest *Hoxa* mRNAs have shorter TLs translated via cap-dependent translation. Since it is much more efficient in developing embryos than cap-independent translation (32), cap-dependent translation would help ensure robust timely expression of these key developmental regulators.

Translational control of Hox genes was first suggested by a report that their translation was reduced in mouse embryos hemizygous for *RPL38* (*Ts/+*)(32). However, the data presented in that study do *not* actually show a decrease in Hox mRNA translation, typically seen as a shift from polysome to monosome and untranslated sucrose gradient fractions. Instead, Hox mRNA were strongly *reduced in*

both polysomes and monosomes in Ts/+ embryos, though the data were presented in separate figures (Figures 3 and 6 from 31). Although the authors reported *Hox* mRNA levels were not decreased in *Ts/+* mutant embryos, the underlying qPCR measurements had such high variance that even considerable changes in mRNA levels would be undetectable. Indeed, qPCR has a number of limitations in estimating mRNA levels (33–35). Additionally, a recent ribosome profiling study in HEK293 cells reported that depletion of *RPL38* decreased the translation efficiency of many genes that promote Hox gene transcription (36). Future ribosome profiling studies from wildtype and Ts/+ embryos are needed to determine whether *RPL38* hemizygosity actually disrupts translation of *Hox* genes and/or their upstream transcriptional regulators.

Our results also do not support the recently reported catalog of 589 hTLs in other mouse genes (16), ninety of which have putative IRESes based on bicistronic reporter assays. We showed these hTLs frequently overlap annotated promoters, enhancers, 3' splice sites, and even protein coding sequences, providing conventional explanations for their unusually high conservation rates. Furthermore, we tested four putative IRES regions from these hTLs and found all encoded promoters. Indeed, the ninety IRES-like hTLs are particularly enriched in annotated promoters and splice sites, compared to non-IRES hTLs, and these features can even be used to predict IRES-like activity from the bicistronic reporter system. This is consistent with our observation that the vast majority of IRES-like hTLs show signs of internal transcription initiation in public RNA-seq from ENCODE and 5' CAGE-seq from RIKEN (28). Taken together, we propose that the vast majority of these hTLs are incorrect due to transcriptome annotation errors and promoter and splicing activities in bicistronic reporter assays. However, the concept of hyperconserved elements in 5' TLs is still intriguing and deserves more careful study to identify genuine hTLs and investigate their functional elements.

It is well known that bicistronic reporter assays are subject to false positive results due to cryptic promoters and splice sites. Thus, several controls are considered essential. These include RNAi treatment to identify monocistronic *Fluc* transcripts, RT-PCR screening for cryptic splicing, and deletion of the SV40 promoter upstream of *Rluc* to account for independent promoter activities (7, 8). Notably, a previous study of Hox gene IRES activity used siRNA targeting *Rluc* as a control for monocistronic transcripts. If only bicistronic transcripts were present, this treatment should deplete both *Rluc* and *Fluc*

mRNA equally. Although siRNA treatment nearly eliminated *Rluc* mRNA, ~30% of *Fluc* mRNA remained, consistent with monocistronic *Fluc* expression driven by promoter activities from the *Hoxa3*, *a4*, *a5*, *a9*, and *a11* IRES-like regions (3). All the putative IRES-like *Hoxa* TLs we tested had independent promoter activity, while non-IRES *Hoxa* TLs did not. Notably, many *Hoxa* TLs also increased spurious expression of the upstream, promoter-less *Rluc* gene (Figure S4). Spurious *Rluc* expression may result from transcriptional enhancers encoded in the misannotated TLs (Figure S1B) and undermines comparisons of *Fluc*/*Rluc* protein and mRNA ratios, previously used to discount cryptic promoters and splicing of putative IRES-like hTLs (16). These observations reinforce the view that bicistronic assays should be replaced by transfection with *in vitro* transcribed mRNA and / or circRNA for studies of cap-independent translation (7). Indeed, recent studies have begun to adopt this more reliable approach, which should allow the accurate identification of genuine cellular IRESes (37).

Our results underscore the importance of accurate transcript annotations for defining and studying TLs. The incorrect, extended *Hoxa9* TL can be traced to experiments that used reverse transcription "primer walking" to find the most upstream 5' end (38). Unfortunately, this appears to have also amplified unspliced introns from *Hoxa9*/*a10* fusion transcripts. Indeed, the region upstream of this misannotated extended TL is extremely G-rich, such that G quadruplexes may have halted reverse transcription. However, such annotation errors go well beyond Hox genes, as many of the extended TLs from other genes also appear to be incorrect (e.g., *Ptp4a1*, *Hoxa4*, *Hoxa7*, etc.). Notably, many of these misannotations overlap introns, suggesting unspliced pre-mRNA may have caused the error. In other cases, transcription initiates at multiple sites within annotated TLs (e.g., *Chrdl1, Dedd*). Astonishingly, even the TL of mouse *Actb* (beta actin) appears to be misannotated in RefGene, initiating with a TATA box and including a promoter (Figure S7). This error may explain why its TL showed apparent IRES-like activity when fused to the *Hoxa9* P4 domain (4). Added to this is the general complexity of mammalian transcriptomes, in which TLs also include introns and 3' splice sites. Together, these issues make accurate mammalian TL annotation particularly challenging, and complicate the study of TL functional elements and conservation. Ongoing efforts to sequence full length transcripts (39, 40), integrated with annotated promoters and transcription start sites, should eventually resolve such issues and greatly aid the study of TL functions in mammals.

Materials and Methods

Luciferase vector cloning

The pRF+423Dux4 plasmid (Addgene #21625) contains Renilla Luciferase (R-luciferase) under the control of a SV40 promoter. Firefly luciferase, downstream of R-luciferase, is transcribed under the control of the same SV40 promoter, and is preceded by a putative upstream IRES (Figure S1). The pRF+423Dux 4 vector was sequenced using primers that anneal to the PGL3 vector (Promega, see primers Table S2). The Dux4 IRES site was deleted from pRF+423Dux4 using PCR primers that flank the IRES region (PRF423DUX4-ATW F and R, Table S2). The primers also incorporated a BglII site after the start codon of F-luciferase, with an upstream HindIII site. The PCR product was phosphatased and circularized by ligation to create the vector pRF-ΔIRES. To delete the SV40 promoter, add an EcoRI site, and remove an additional BglII site, pRF-ΔIRES was used as a template for a second PCR, using the primers SV40D-EcoRI and SV40D-XhoI-R (Table S2). The resulting PCR product was phosphatased and circularized by ligation to create the vector pRF-ΔSV40. Both pRF-ΔIRES and pRF-ΔSV40 were verified by Sanger sequencing and tested for luciferase activity in C3H/10T1/2 mouse embryonic fibroblast cells (MEF) obtained from ATCC. A R-luciferase only vector was constructed by removing F-luciferase from pRF-ΔIRES by XbaI digestion. The XbaI cut vector was gel purified and circularized by ligation, resulting in the vector pR+SV40. The pR+SV40 vector was Sanger sequenced and its luciferase activity was verified in MEF cells (data not shown).

Putative 5'UTR cloning

Putative 5'UTR sequences were obtained as double stranded DNA fragments from Twist Biosciences and Genewiz (Table S2). The DNA fragments were PCR amplified, digested with HindIII and BglII, and cloned into the pRF-ΔSV40 vector at the HindIII and BglII sites upstream of F-luciferase. Due to limitations in DNA synthesis, five additional As were added by site-directed mutagenesis using MMHOXA11-ATW forward and reverse primers (Table S2) to finish the *Hoxa11* construct. Additional sequences in the *Hoxa3* hTL were removed by PCR using primers HOXA3-SATW forward and reverse primers (Table S2). Site-directed mutagenesis was also performed on Hoxa3, a5, a7, a9, and a11 constructs to mutate E box

sites to CACTAT. Some inserts affected both F-luciferase and R-luciferase (Figure S4), For more precise ratiometric measurements, the R-luciferase gene was removed from wildtype and E-box mutant *Hoxa3*, *a5*, *a7*, *a11*, and *a9* constructs by EcoRI and HindIII digestion, end polishing with DNA polymerase I large fragment (Klenow), and re-ligation (pF-ΔSV40; Figure 3). All constructs were Sanger sequenced (Table S2), and transfection grade plasmid DNA was purified using a Qiagen Plasmid Mini column according to the manufacturer's instructions.

Luciferase Assays

In 96-well tissue culture plates, 2 x 10^3 mouse embryonic fibroblast (C3H/10T1/2 clone 8, ATTC) cells seeded in 100 µl DMEM supplemented with 10% FBS per well. Cells were allowed to adhere and grow for 24 hours at 37 °C. In 10 µl of Opti-MEM™. 100 ng of construct was mixed with 0.4 µl of ViaFect™(Promega) and incubated for 12 minutes at room temperature. The transfection mixture was added dropwise to the cells and the cells were incubated at 37 °C for 24 hours. F-luciferase and R-luciferase expression was assayed in a TECAN Spark plate reader using the Dual-Glo® Luciferase Assay System (Promega) according to the manufacturer's instructions. Both F-luciferase and R-luciferase were measured for 10 seconds per well. Each construct was transfected in triplicate.

USF1 and 2 siRNA knockdown

In a 96 well tissue culture plate, 1 x 10^3 mouse embryonic fibroblast (C3H/10T1/2 clone 8, ATTC) cells seeded in 100 µl DMEM supplemented with 10% FBS per well. Cells were allowed to adhere and grow for 24 hours at 37 °C. In 10 µl of Opti-MEM™, 1 pmol of siRNAs (scrambled control or USF1/2 siRNA; Santa Cruz Biotechnologies) were mixed with 0.3 µl of Lipofectamine™ 3000 transfection agent and incubated at room temperature for 15 minutes. The transfection mixture was added dropwise to the cells so that the final concentration of each siRNA was 10 nM. The cells were incubated at 37 °C for 24 hours. For each well, 20 ng of pR+SV40 vector (Renilla only) and 80 ng of a Hox gene construct were mixed with 10 µl Opti-MEM™ and 0.5 µl of ViaFect™ and incubated for 15 minutes at room temperature. The mixture was added dropwise to the cells and the cells were incubated at 37 °C for 24 hours. F-luciferase and R-

luciferase were assayed as described above (Luciferase Assay). Each construct was transfected in triplicate.

Validation of USF1 and USF2 siRNA knockdown

In a 6 well tissue culture plate, 3 x 10^4 C3H/10T1/2 cells seeded in 2 ml DMEM supplemented with 10% FBS per well. Cells were allowed to adhere and grow for 24 hours at 37 °C. In 250µl of Opti-MEM™, 20 pmol of siRNAs (either scrambled control or USF1/USF2 siRNA) were mixed with 7.5 µl of Lipofectamine™ 3000 transfection agent and incubated at room temperature for 15 minutes. The transfection mixture was added dropwise to the cells so that the final concentration of each siRNA was 10 nM. The cells were incubated at 37 °C for 48 hours. The media were removed, and total RNA was extracted using TRIzol™ reagent. RNA was twice treated with TURBO™ DNase and RT-qPCR was performed using the SuperScript™ III Platinum™ SYBR™ Green One-Step RT-qPCR kit according to the manufacturer's instructions (see Table S2 for primer sequences). Relative gene expression of USF1 and USF2 were compared to GAPDH using the ΔΔ-Ct method. Three technical replicates were performed for each gene, along with three technical replicates of no template controls. No amplification was detected for the no template controls.

Logistic Regression

We used active and non-active hTLs provided by Gun Woo Byeon and Maria Barna (16). After removing records from TLs that were not previously classified as hyperconserved, the dataset included 133 nonactive hTLs and 90 active hTLs. We compiled a list of several categorical and numerical sequence features that could contribute to bicistronic activity (e.g. GC content, CAGE data, ebox motifs; Figure 4). Transcript leaders lacking sufficient data for refTSS calculations (refTSSreads5P) were assigned a mean imputation filler value. To perform classification of active versus non-active transcript leaders, we used LogisticRegressionCV from scikit learn (sklearn.linear_model.LogisticRegressionCV) with the default solver=lbfgs, Cs = 10, intercept=True, and cv = 10 parameters. All numerical features were normalized using sklearn.preprocessing.MinMaxScaler. 100 separate models were individually trained on random samples of 80% of the data and tested on the remaining 20% of the data.

ENCODE RNA-Seq Data

From the ENCODE database, we used polyA plus RNA-seq data from *Mus. musculus* and *Homo sapiens* tissues that showed evidence of transcription for *Hoxa9*, assessed by visual examination. For the tissues containing multiple bigwig files, we merged the reads to create a new complete bigwig. The files containing "negative strand signal" data were used, because *Hoxa9* is on the negative strand. If no negative specific file was given, then the "all reads" signal was used. For positive strand gene examples, the "positive strand signal" files were used instead. The accession numbers used are included in Table S3.

CAGE data (RefTSS)

CAGE-seq data were downloaded using SRA Run Selector from NCBI from SRA study number DRP000949 (BioProject PRJDB1980). In this study CAGE reads were obtained from Human and Mouse transcripts to define transcription start sites (TSSs). For our study, we used the *Mus musculus* data from runs DRR003905 (experiment DRX003141) and DRR003906 (experiment DRX003142). The data were from induced pluripotent stem cells (iPCSs) and embryonic stem cells (ESs) respectively. Reads were processed using fastq-dump followed by cutadapt. The processed data was aligned to the mouse genome using STAR. Reads were summed and assigned to annotated refTSS peaks via bedtools intersect to define refTSS strength. Files used are listed in Table S3.

GWIPS RNA-Seq Data

RNA-seq data was retrieved from GWIPS table browser(41) from Mouse (mm10) using the global aggregate setting. Data was compiled from 26 files (listed below). Bedgraphs were combined using bedtools unionbedg.

Alvarez Dominguez 2017, Atger 2015, Atlasi 2020, Blanco 2016, Castaneda 2014, Castela Szekely 2017, Cho 2015, Eichhorn 2014, Fradejas-Villar 2016, Freimer 2018, Gonzalez 2014, Guo 2010, Hornstein 2016, Howard 2013, Hurt 2013, Ingolia 2011, Janich 2015, Laguesse 2015, Rapino 2018, Reid 2014, Reid 2016, Reid 2017, Sendoel 2017, Simsek 2017, Thoreen 2012, You 2015

Infernal & Rscape

The sequence for the predicted mouse Hoxa9 IRES from (3) was used. Using the latest Zoonomia Cactus mammalian genome alignment (17), we mapped the Mouse coordinates (chr6:52226238-52226413) to 208 mammalian species via halLiftover (42). An additional 23 vertebrate sequences were extracted from the UCSC database for a total of 231 species. The sequences and results are in Table S1. The conserved IRES structure from (3) was converted into dot-bracket notation and used to generate a Stockholm format file containing the 231 sequences, the conserved structure, and the *Mus musculus* sequence as reference.

The Infernal package was used to build a covariance model and prune the sequence alignment. Using default parameters for cmbuild and cmcalibrate, 25 close species were used to build and calibrate the initial covariance model. Using cmsearch, target sequences (230 sequences) with appropriate E-values (default) for covariation testing were kept (medaka not significant, filtered out). A new Stockholm file was generated from the remaining sequences. The resulting file was used as input to R-scape using default parameters with various E-value thresholds (0.005, 0.01, 0.1, 1, 10, 20, 30). The IRES data as well as the null files were tested seven times with the varying E-values. The results showed that the IRES sequences showed no significant covariation and had less covariation than the random mutation null models tested.

Acknowledgments

We would like to thank Cassia Williams-Rogers, Alyssa Lawler, and Irene Kaplow for assistance in computational data analysis and John Woolford for helpful discussions and anonymous reviewers for suggestions to improve the manuscript. We thank Gun Woo Byeon and Maria Barna for providing a table of hTL bicistronic assay data previously summarized in (16). This study was supported by NIH R01 GM121895 to CJM.

References

1. A. G. Hinnebusch, I. P. Ivanov, N. Sonenberg, Translational control by 5'-untranslated regions of eukaryotic mRNAs. *Science* **352**, 1413–1416 (2016).

2. J. Pelletier, N. Sonenberg, The Organizing Principles of Eukaryotic Ribosome Recruitment. *Annu. Rev. Biochem.* **88**, 307–335 (2019).

3. S. Xue, *et al.*, RNA regulons in Hox 5' UTRs confer ribosome specificity to gene regulation. *Nature* **517**, 33–38 (2015).

4. K. Leppek, *et al.*, Gene- and Species-Specific Hox mRNA Translation by Ribosome Expansion Segments. *Mol. Cell* **80**, 980–995.e13 (2020).

5. K. Leppek, G. W. Byeon, K. Fujii, M. Barna, VELCRO-IP RNA-seq reveals ribosome expansion segment function in translation genome-wide. *Cell Rep.* **34**, 108629 (2021).

6. M. Kozak, A second look at cellular mRNA sequences said to function as internal ribosome entry sites. *Nucleic Acids Res.* **33**, 6593–6602 (2005).

7. I. M. Terenin, V. V. Smirnova, D. E. Andreev, S. E. Dmitriev, I. N. Shatsky, A researcher's guide to the galaxy of IRESs. *Cell. Mol. Life Sci.* **74**, 1431–1455 (2017).

8. Y. Yang, Z. Wang, IRES-mediated cap-independent translation, a path leading to hidden proteome. *J. Mol. Cell Biol.* **11**, 911–919 (2019).

9. B. Han, J.-T. Zhang, Regulation of gene expression by internal ribosome entry sites or cryptic promoters: the eIF4G story. *Mol. Cell. Biol.* **22**, 7372–7384 (2002).

10. B. Han, *et al.*, Regulation of constitutive expression of mouse PTEN by the 5'-untranslated region. *Oncogene* **22**, 5325–5337 (2003).

11. M. E. Van Eden, M. P. Byrd, K. W. Sherrill, R. E. Lloyd, Demonstrating internal ribosome entry sites in eukaryotic mRNAs using stringent RNA test procedures. *RNA* **10**, 720–730 (2004).

12. B. T. Baranick, *et al.*, Splicing mediates the activity of four putative cellular internal ribosome entry sites. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 4733–4738 (2008).

13. A. G. Bert, R. Grépin, M. A. Vadas, G. J. Goodall, Assessing IRES activity in the HIF-1alpha and other cellular 5' UTRs. *RNA* **12**, 1074–1083 (2006).

14. R. J. Jackson, The current status of vertebrate cellular mRNA IRESs. *Cold Spring Harb. Perspect. Biol.* **5**, a011569–a011569 (2013).

15. S. Weingarten-Gabbay, *et al.*, Comparative genetics. Systematic discovery of cap-independent translation sequences in human and viral genomes. *Science* **351** (2016).

16. G. W. Byeon, *et al.*, Functional and structural basis of extreme conservation in vertebrate 5' untranslated regions. *Nat. Genet.* **53**, 729–741 (2021).

17. Zoonomia Consortium, A comparative genomics multitool for scientific discovery and conservation. *Nature* **587**, 240–245 (2020).

18. E. P. Nawrocki, S. R. Eddy, Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).

19. E. Rivas, J. Clements, S. R. Eddy, A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nat. Methods* **14**, 45–48 (2017).

20. E. Rivas, J. Clements, S. R. Eddy, Estimating the power of sequence covariation for detecting conserved RNA structure. *Bioinformatics* **36**, 3072–3076 (2020).

21. ENCODE Project Consortium, *et al.*, Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).

22. R. Dreos, G. Ambrosini, R. C. Périer, P. Bucher, The Eukaryotic Promoter Database: expansion of EPDnew and new promoter analysis tools. *Nucleic Acids Res.* **43**, D92–6 (2015).

23. C. E. Grant, T. L. Bailey, W. S. Noble, FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).

24. L. Desbarats, S. Gaubatz, M. Eilers, Discrimination between different E-box-binding proteins at an endogenous target gene of c-myc. *Genes Dev.* **10**, 447–460 (1996).

25. S. Corre, M.-D. Galibert, Upstream stimulating factors: highly versatile stress-responsive transcription factors. *Pigment Cell Res.* **18**, 337–348 (2005).

26. X. de Martin, R. Sodaei, G. Santpere, Mechanisms of Binding Specificity among bHLH Transcription Factors. *Int. J. Mol. Sci.* **22** (2021).

27. H. Zhang, *et al.*, Functional Interrogation of HOXA9 Regulome in MLLr Leukemia via Reporter-based CRISPR/Cas9 screen (2020) https:/doi.org/10.1101/2020.04.20.050583 (October 17, 2021).

28. I. Abugessaisa, *et al.*, refTSS: A Reference Data Set for Human and Mouse Transcription Start Sites. *J. Mol. Biol.* **431**, 2407–2422 (2019).

29. S. Lin, *et al.*, Comparison of the transcriptional landscapes between human and mouse tissues. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 17224–17229 (2014).

30. I. P. Ivanov, J. A. Saba, C. Fan, J. Wang, A. E. Firth, C. Cao, R. Green, T.E. Dever, Evolutionarily Conserved Inhibitory uORFs Sensitize *Hox* mRNA Translation to Start Codon Selection Stringency. *PNAS* in press (2022).

31. A. R. Gruber, R. Lorenz, S. H. Bernhart, R. Neuböck, I. L. Hofacker, The Vienna RNA websuite. *Nucleic Acids Res.* **36**, 70–74 (2008).

32. N. Kondrashov, *et al.*, Ribosome-mediated specificity in Hox mRNA translation and vertebrate tissue patterning. *Cell* **145**, 383–397 (2011).

33. F. Souazé, A. Ntodou-Thomé, C. Y. Tran, W. Rostène, P. Forgez, Quantitative RT-PCR: limits and accuracy. *Biotechniques* **21**, 280–285 (1996).

34. S. A. Bustin, T. Nolan, Pitfalls of quantitative real-time reverse-transcription polymerase chain reaction. *J. Biomol. Tech.* **15**, 155–166 (2004).

35. L. Negură, A. Negură, Limits and Pitfalls of Sybr Green Detection in Quantitative Pcr. *Secţiunea Genetică şi Biologie Moleculară* **13**, 13–18 (2012).

36. A. V. Gopanenko, *et al.*, Knockdown of the Ribosomal Protein eL38 in HEK293 Cells Changes the Translational Efficiency of Specific Genes. *Int. J. Mol. Sci.* **22** (2021).

37. C.-K. Chen, *et al.*, Structured elements drive extensive circular RNA translation. *Mol. Cell* **81**, 4300–4318.e13 (2021).

38. S. Fujimoto, *et al.*, Analysis of the murine Hoxa-9 cDNA: an alternatively spliced transcript encodes a truncated protein lacking the homeodomain. *Gene* **209**, 77–85 (1998).

39. E. Rebboah, *et al.*, Mapping and modeling the genomic basis of differential RNA isoform expression at single-cell resolution with LR-Split-seq. *Genome Biol.* **22**, 286 (2021).

40. S. Y. Anvar, *et al.*, Full-length mRNA sequencing uncovers a widespread coupling between transcription initiation and mRNA processing. *Genome Biol.* **19**, 46 (2018).

41. A. M. Michel, *et al.*, GWIPS-viz: development of a ribo-seq genome browser. *Nucleic Acids Res.* **42**, D859–64 (2014).

42. X. Zhang, I. M. Kaplow, M. Wirthlin, T. Y. Park, A. R. Pfenning, HALPER facilitates the identification of regulatory element orthologs across species. *Bioinformatics* **36**, 15 (2020).

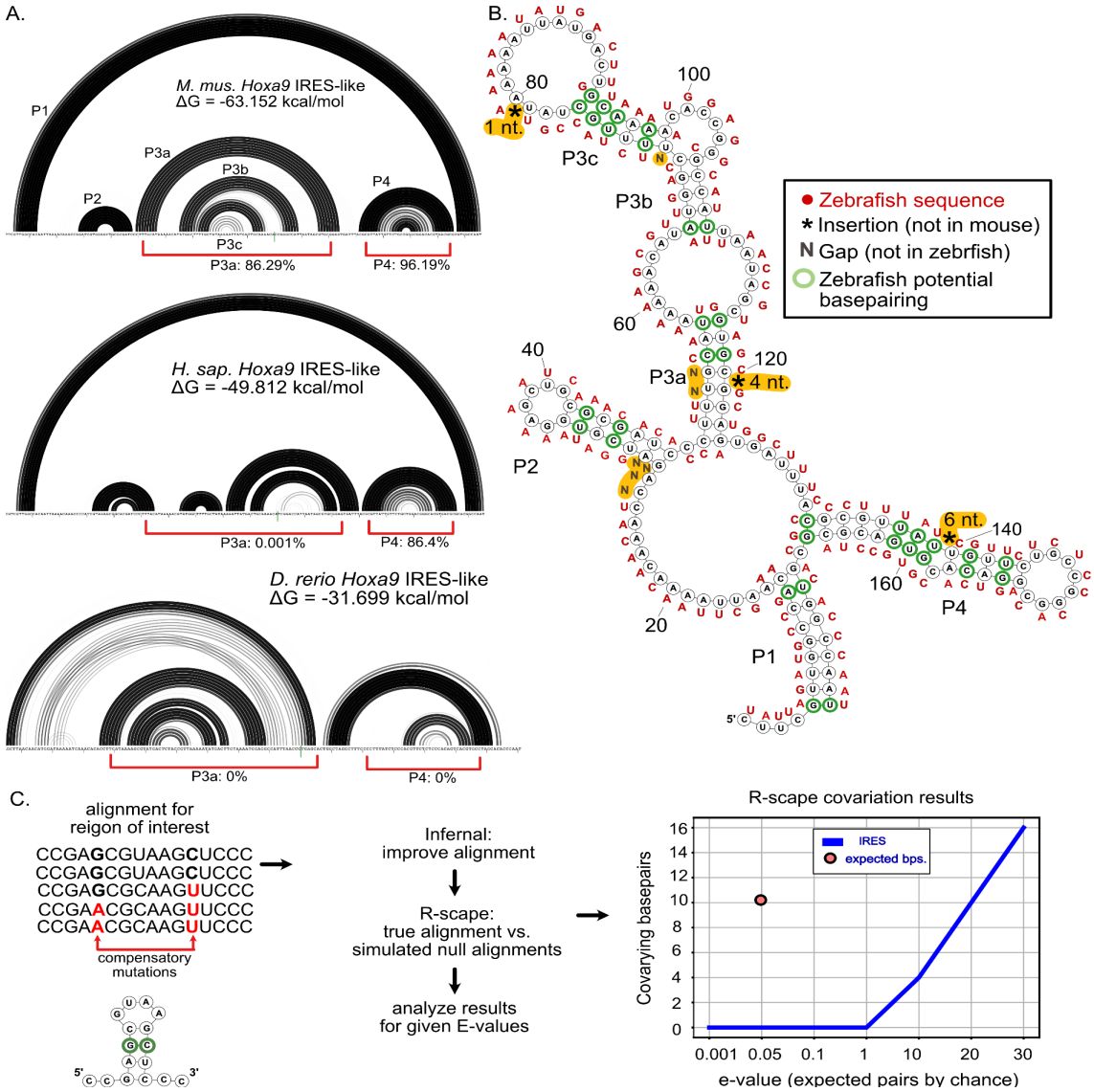


Figure 1. The proposed structure of the putative Hoxa9 IRES element is not conserved. (A) Rainbow graphs depict the probability of base pairing for the Hoxa9 IRES-like region from mouse (top), human (middle) and zebrafish (bottom). Pairing regions (stems) are numbered as in (3). Base pairing probabilities were determined using RNAstructure. The mouse model is highly consistent with the published model(3). Red brackets indicate the frequency of P3a and P4 helix formation in 10,000 predicted suboptimal structures. Human and mouse Hoxa9 share P1 and P4, but lack P3, which was reported to be essential for IRES activity (3). Zebrafish Hoxa9 does not share any structural similarity with mammalian homologs, despite driving bicistronic reporter activity (3). (B) Secondary structure model of mouse Hoxa9 putative IRES region (3). Corresponding zebrafish sequences are shown in red. Most proposed base pairs are not conserved. Zebrafish has insertions (asterices) and deletions (Ns) in the critical P3 and P4 elements. (C) Results of R-scape analysis of mutual information for the mouse Hoxa9 putative IRES region using alignments from 208 mammals and 23 other vertebrates. The number of covarying sites (y axis) is given for different e-value cutoffs (x axis). Although the alignment has the power to detect ~ 10 compensatory pairs (red point, Table S1), covarying base pairs are less common than expected by chance in the IRES-like element (blue line).

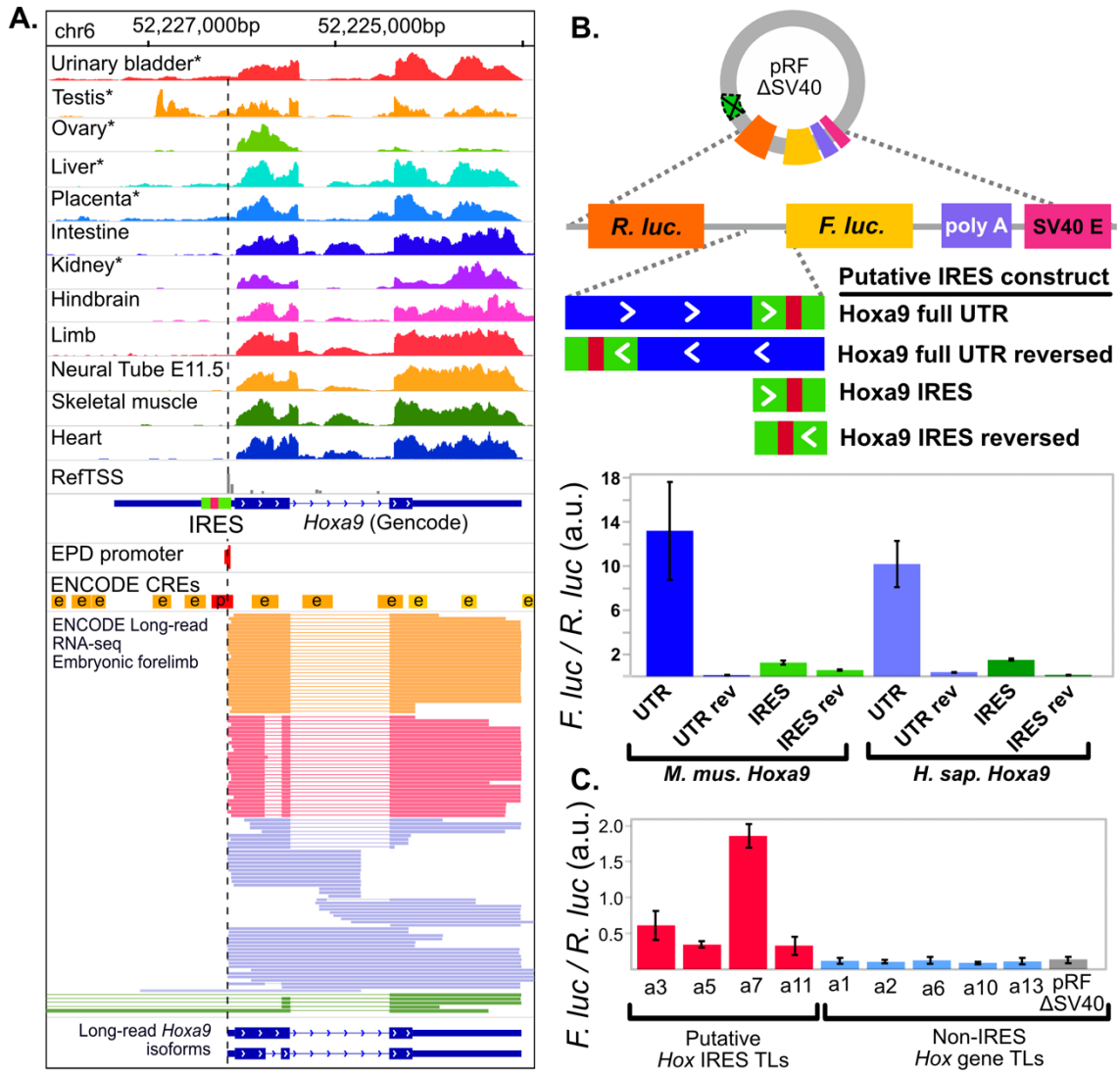


Figure 2. The putative IRES-like domains of *Hoxa9* and other *Hox* genes encode functional promoters. (A) The IRES-like region of mouse *Hoxa9* is not transcribed. ENCODE short- (upper) and long-read (PacBio, lower) RNA-seq data support transcription initiation downstream of the putative IRES-like domain in most mouse tissues, coinciding with an annotated promoter from EPD and ENCODE and 5' CAGE data and refTSS sites. Long-read data additionally suggest the extended isoform annotation may reflect rare unspliced RNA from *Hoxa9/a10* fusion transcripts (green). Asterisks denote strand-specific RNA-seq (B) The putative *Hoxa9* IRES is a promoter. The SV40 promoter was deleted from the pRF bicistronic vector. Putative IRES regions were cloned between *Renilla* and *Firefly* luciferase and tested for activity in C3H10T1/2 cells. Bar graphs show the *F. luc* / *R. luc* ratio indicating promoter activity from mouse and human *Hoxa9* regions. The extended UTR and IRES-like regions function as independent promoters in the forward orientation. (C) Putative IRES-like regions from other mouse *Hoxa* genes function as promoters. Annotated transcript leaders from each *Hoxa* gene were tested as in (B). TLs containing putative IRES-like elements drove expression, while non-IRES TLs had background expression levels. Error bars show 95% confidence intervals with n = 3.

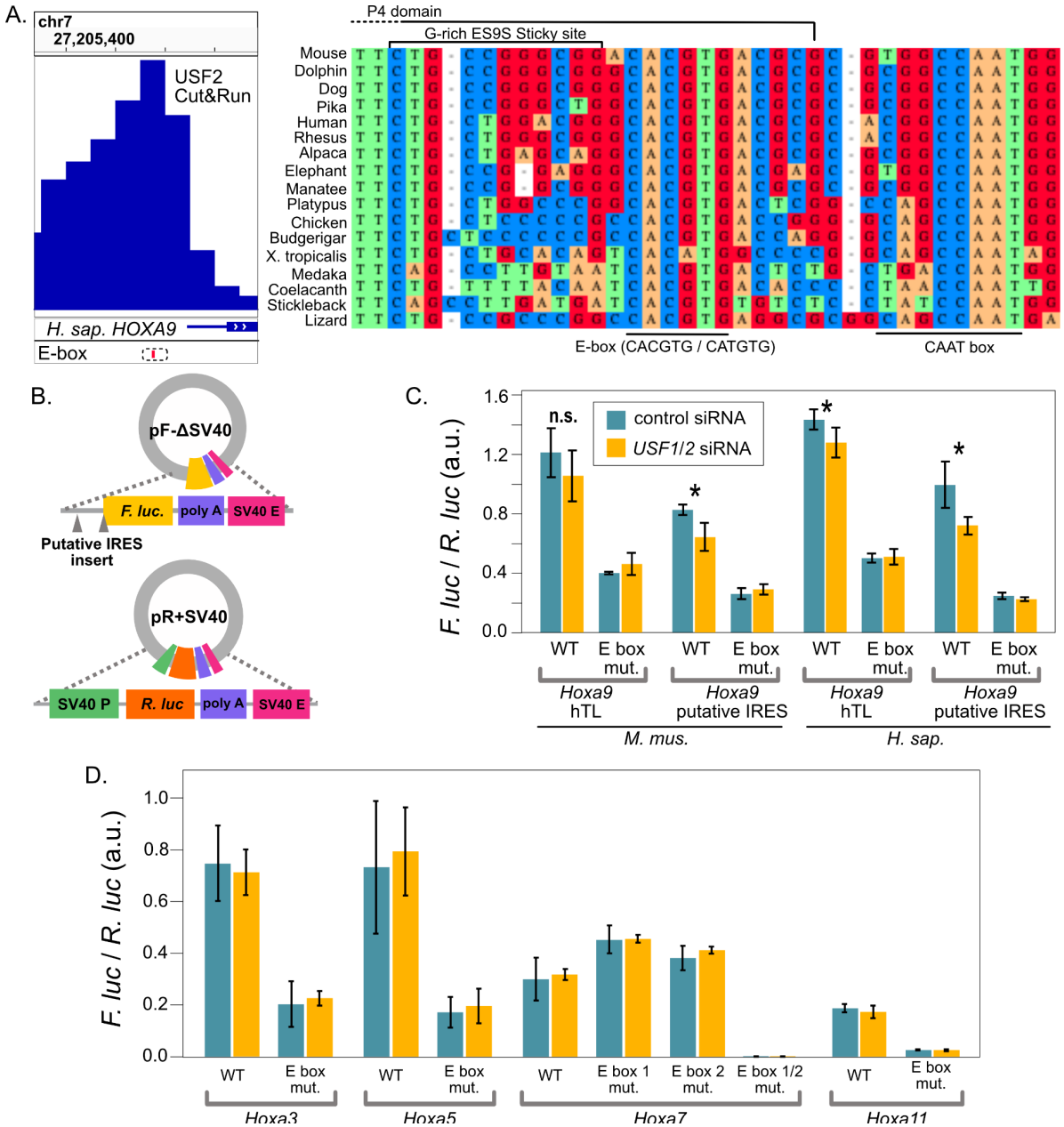
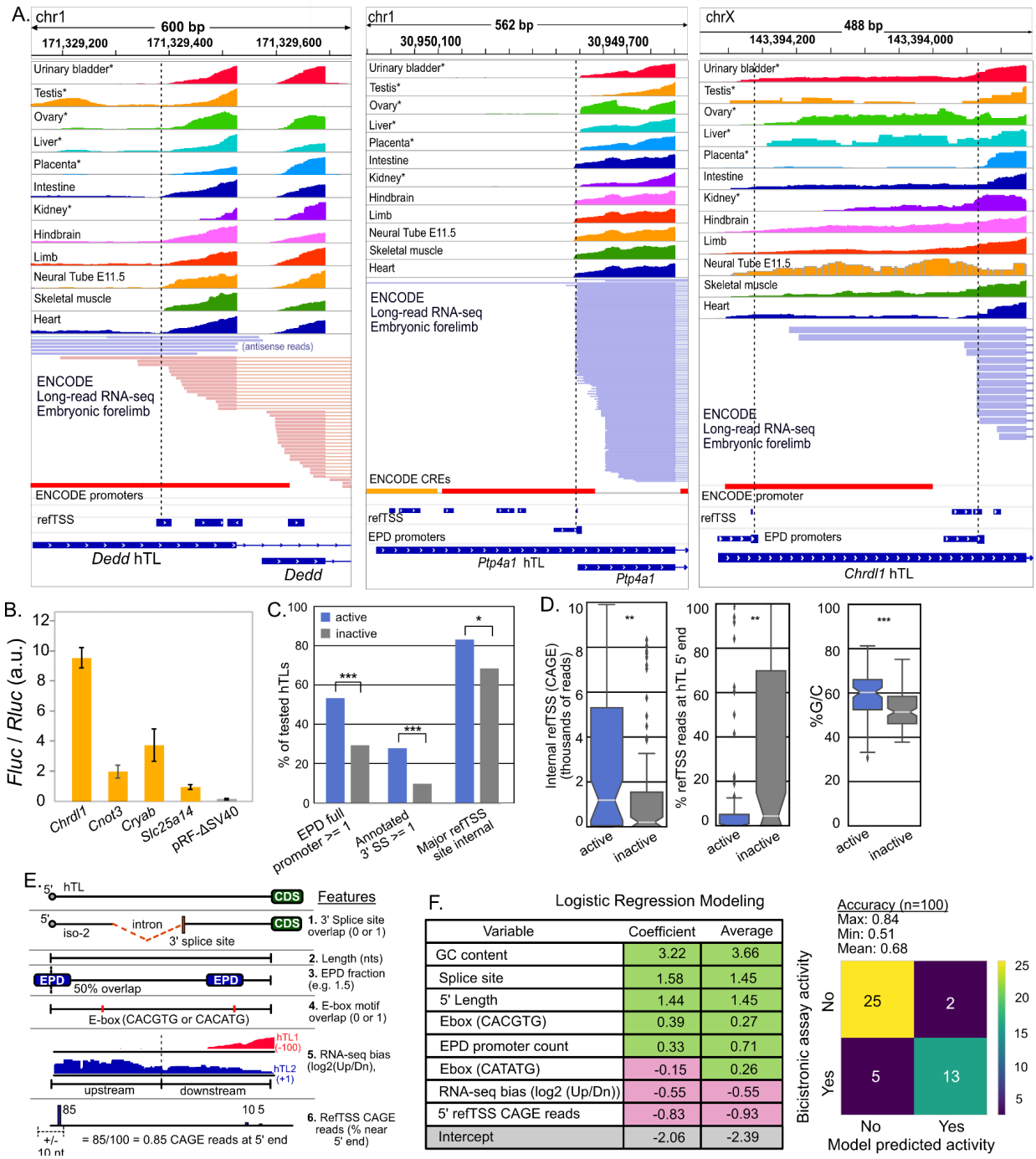


Figure 3. **Putative IRES-like elements in *Hoxa* genes contain functional E-boxes recognized by USF1/2.** (A) USF2 binds at a conserved E-box upstream of human *Hoxa9*. Genome browser view of CUT&RUN data show human *USF2* binding upstream of *Hoxa9* (ref. 27. The dashed box shows the location of a hyperconserved E-box motif. The alignment shows sequences from the homologous regions of diverse representative vertebrate genomes upstream of *Hoxa9*. The P4 domain sequences and the non-conserved G-rich ES9S interaction site are shown above, while the hyperconserved E-box and CAAT-box are noted below. (B) The *F. luc.* and *R. luc.* reporter genes were put on independent plasmids to test the functions of E-box elements. (C) Mutations of the E-box motifs from mouse and human *Hoxa9* decreased expression in C3H10T1/2 cells. siRNA co-depletion of USF1 and USF2 (Fig. S8) further decreased expression (one-tail Welch's t-test, * $P < 0.05$), but did not affect the expression of E-box mutant reporters. (D) Mutation of E-boxes from putative IRESes of *Hoxa3*, *a5*, a7, and *a11* decreased promoter activity. Single E-box mutations show slight increases in *Hoxa7* promoter activity, while the double mutation eliminated promoter function. Error bars show 95% confidence intervals with n = 3.



Variable	Coefficient	Average
GC content	3.22	3.66
Splice site	1.58	1.45
5' Length	1.44	1.45
Ebox (CACGTG)	0.39	0.27
EPD promoter count	0.33	0.71
Ebox (CATATG)	-0.15	0.26
RNA-seq bias (log2 (Up/Dn))	-0.55	-0.55
5' refTSS CAGE reads	-0.83	-0.93
Intercept	-2.06	-2.39

Figure 4. Putative IRES-like hTLs can be explained by promoters and 3' splice sites due to 5' UTR annotation errors. (A) Examples of promoter overlap commonly seen in putative IRES-like hTLs. Short-read (upper) and long-read (lower) RNA-seq data show transcription often initiates internally, coinciding with annotated promoters (ENCODE and EPD) and transcription start sites (refTSS). (B) The hTLs from four putative mouse IRESes have promoter activity in pRF-ΔSV40 transfected C3H10T1/2 cells. Error bars depict 95% confidence intervals with n = 3. (C) hTLs with IRES-like activity are enriched in EPD promoters, 3' splice sites, and major internal TSS sites (X^2 tests). (D) IRES-active hTLs have significantly more internal CAGE 5' reads, a lower fraction of TSS reads at annotated 5' ends and higher G/C content than IRES-inactive hTLs (Wilcoxon tests). (E) Features for logistic regression modeling. RNA-seq bias is the ratio of reads in upstream and downstream hTL halves across GWIPs-Viz (41) RNA-seq datasets. RefTSS CAGE reads are the percentage of 5' end reads mapped near the annotated transcription start site (data from 28). (F) Logistic regression modeling of IRES-like and non-IRES hTLs. Features associated with internal promoters (GC content, EPD promoter count fraction, E-boxes) and splice sites are positively correlated with

bicistronic reporter expression, while features associated with full length TLs (CAGE reads at annotated 5' ends and RNA-seq 5' end bias) are negatively correlated with bicistronic reporter activity. 100 models were generated, with an average accuracy of 68%. *, **, and *** denote $P < 0.05$, 0.01, and 0.001, respectively.

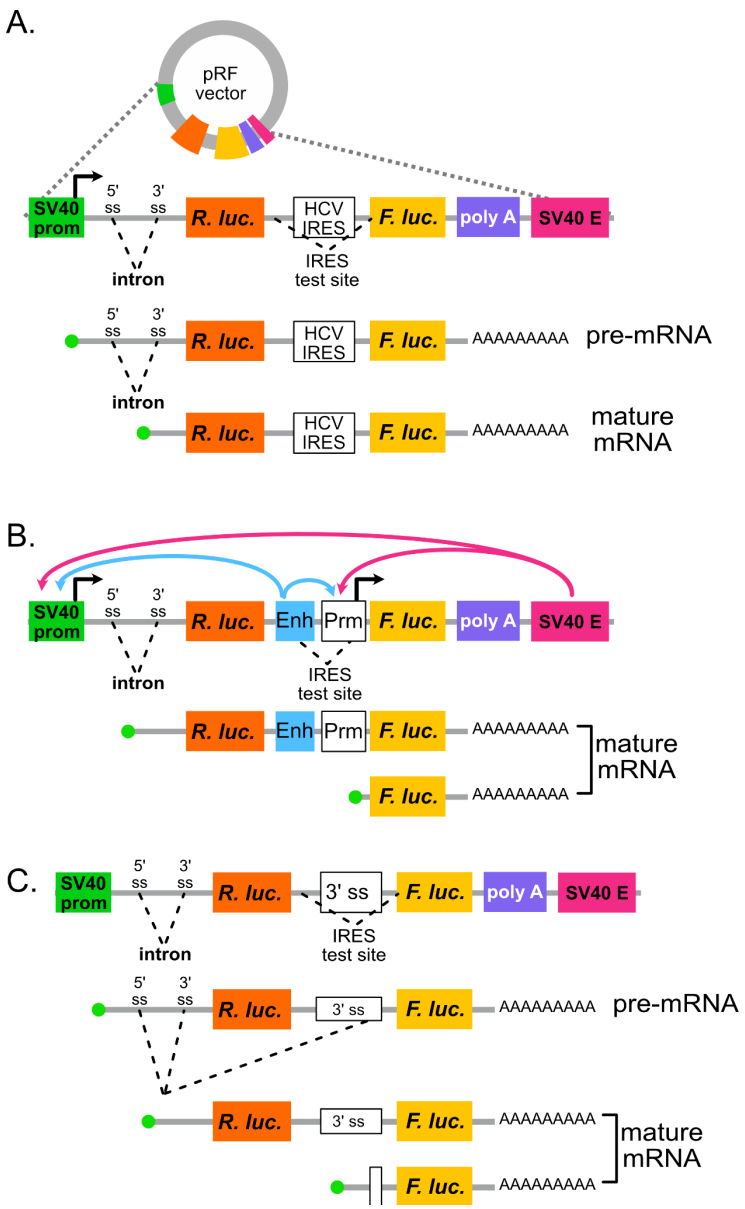


Figure S1. The classical pRF vector / bicistronic reporter is subject to cryptic promoter and splicing activity. (A) Design of the pRF vector and expected mRNA products. Transcription is driven by the SV40 promoter (green), increased by an included SV40 enhancer sequence (magenta). An intron is positioned upstream of Renilla luciferase to increase expression by splicing-mediated mRNA export. An IRES test site is positioned between Renilla and Firefly luciferase ORFs, and is shown with the Hepatitis C Virus IRES (HCV). Transcription and splicing are expected to generate homogenous bicistronic transcripts. (B) Enhancer and promoter sequences placed in the IRES test site can generate monocistronic Firefly luciferase transcripts that give a false-positive "IRES" signal. The "IRES" enhancer sequence (blue) can alter the transcription of both bicistronic and monocistronic transcripts. (C) The presence of 3' splice sites (both strong and cryptic) in the IRES test site can generate monocistronic transcripts by splicing to the upstream intronic 5' splice site. This can also create a false-positive "IRES" signal. Removal of the canonical intron can reduce splicing, but cryptic 5' splice sites can still lead to trace levels of Firefly luciferase expression.

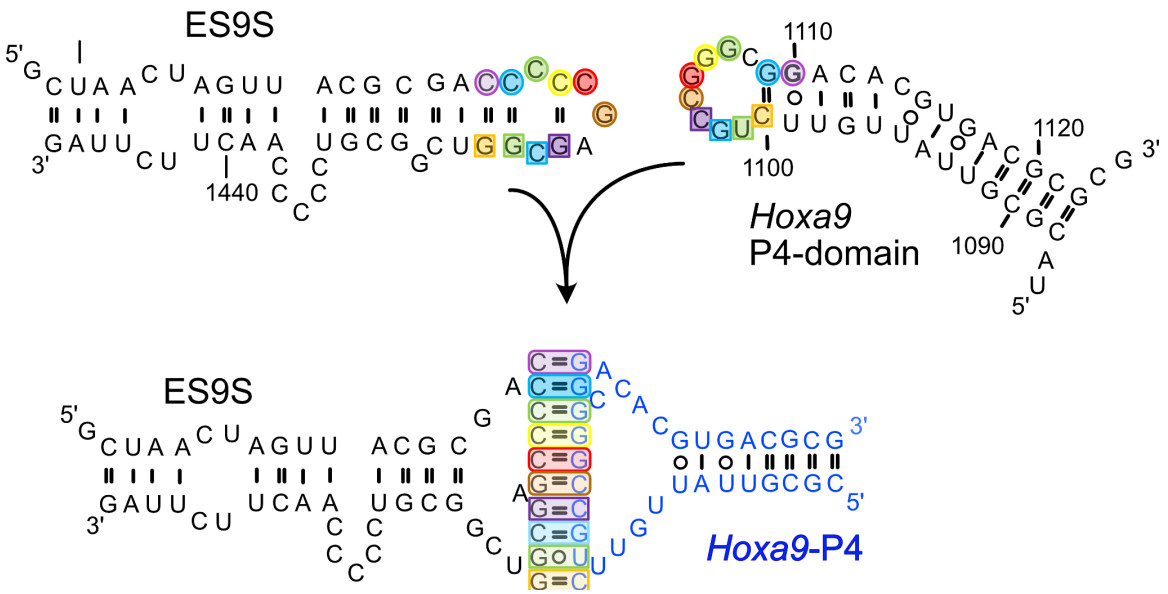


Figure S2. Mouse ES9S and putative *Hoxa9* IRES P4-domain have complementary sequences that could support a kissing stem loop interaction. Proposed structures of individual RNAs are depicted as reported from Leppek et al., 2020 and Xue et al., 2015 (above). Nucleotides that have the propensity to pair are shaded in matched colored ovals and squares. A G-rich segment in the putative *Hoxa9* IRES P4-domain is complementary to a C-rich segment in ES9S, with further potential pairing between additional adjacent nucleotides.

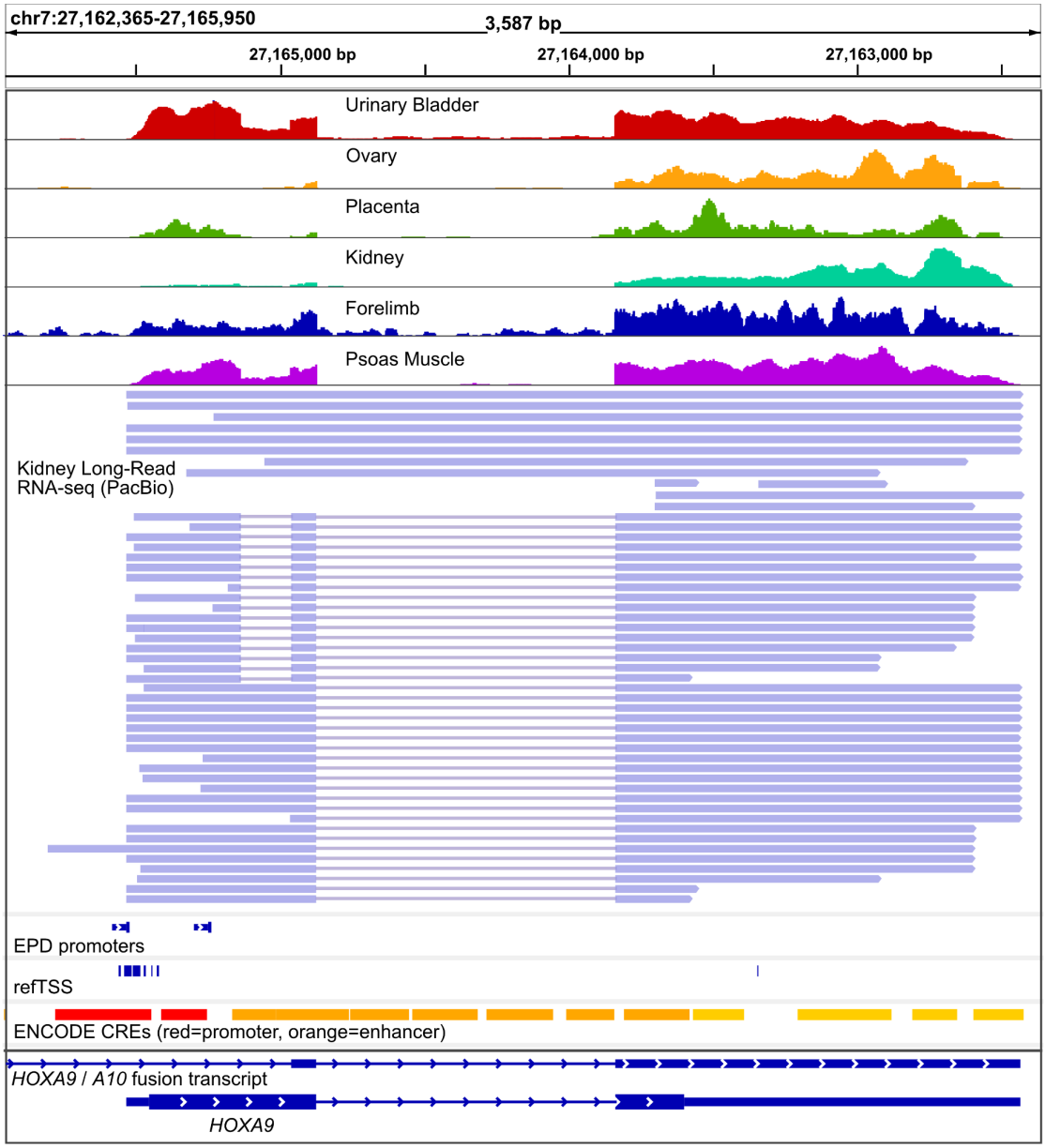


Figure S3. Human *Hoxa9* expresses short 5' UTR isoforms excluding the putative IRES. Genome browser tracks show short read polyA RNA-seq data from the ENCODE project consortium from a variety of representative human tissues (upper, rainbow colors). Long-read (PacBio) RNA-seq data from the ENCODE project (lower) shows two predominant isoforms whose transcripts initiate close to the *Hoxa9* protein coding sequence, consistent with annotated promoters (EPD and ENCODE) and transcription start sites (refTSS). The human *Hoxa9* gene annotation also includes a *Hoxa9*/*Hoxa10* fusion transcript, consistent with the fusion transcript seen in mouse long-read RNA-seq data (Figure 2).

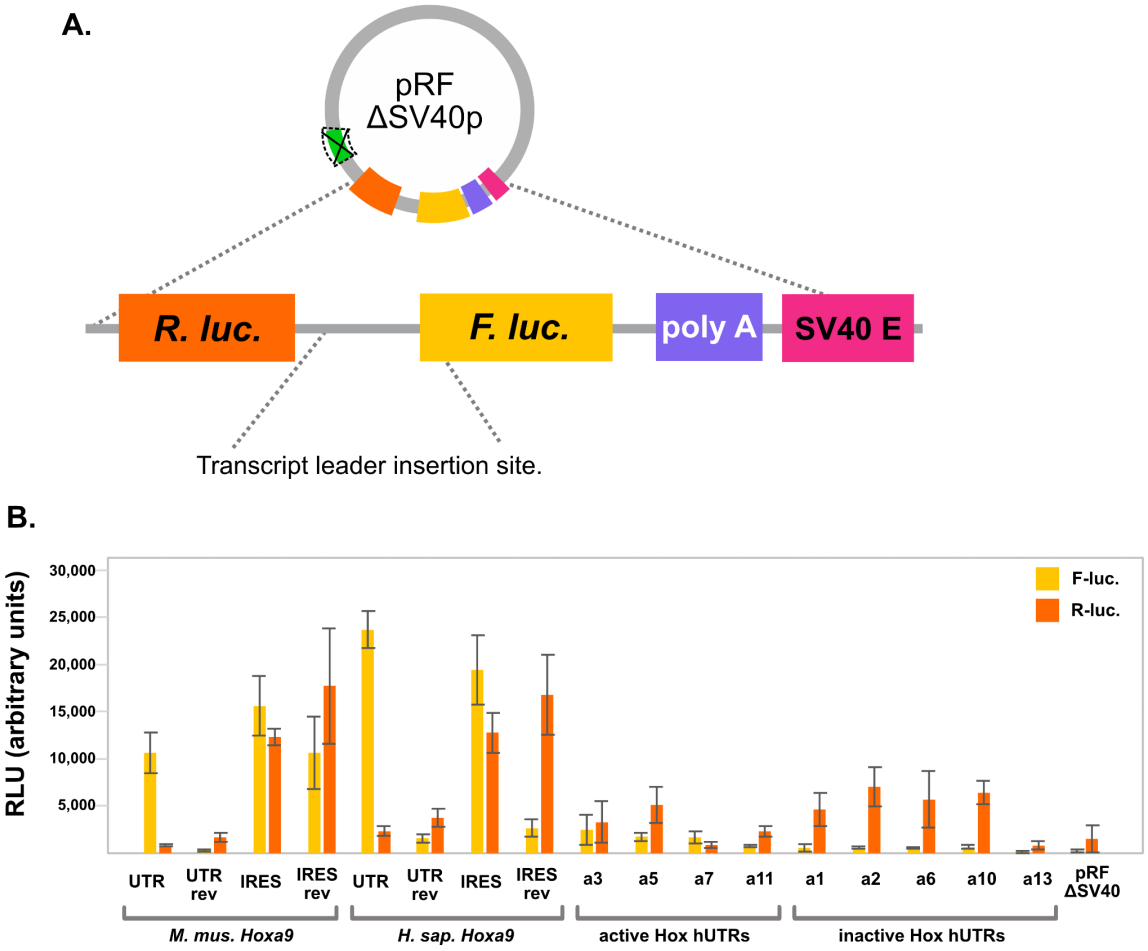
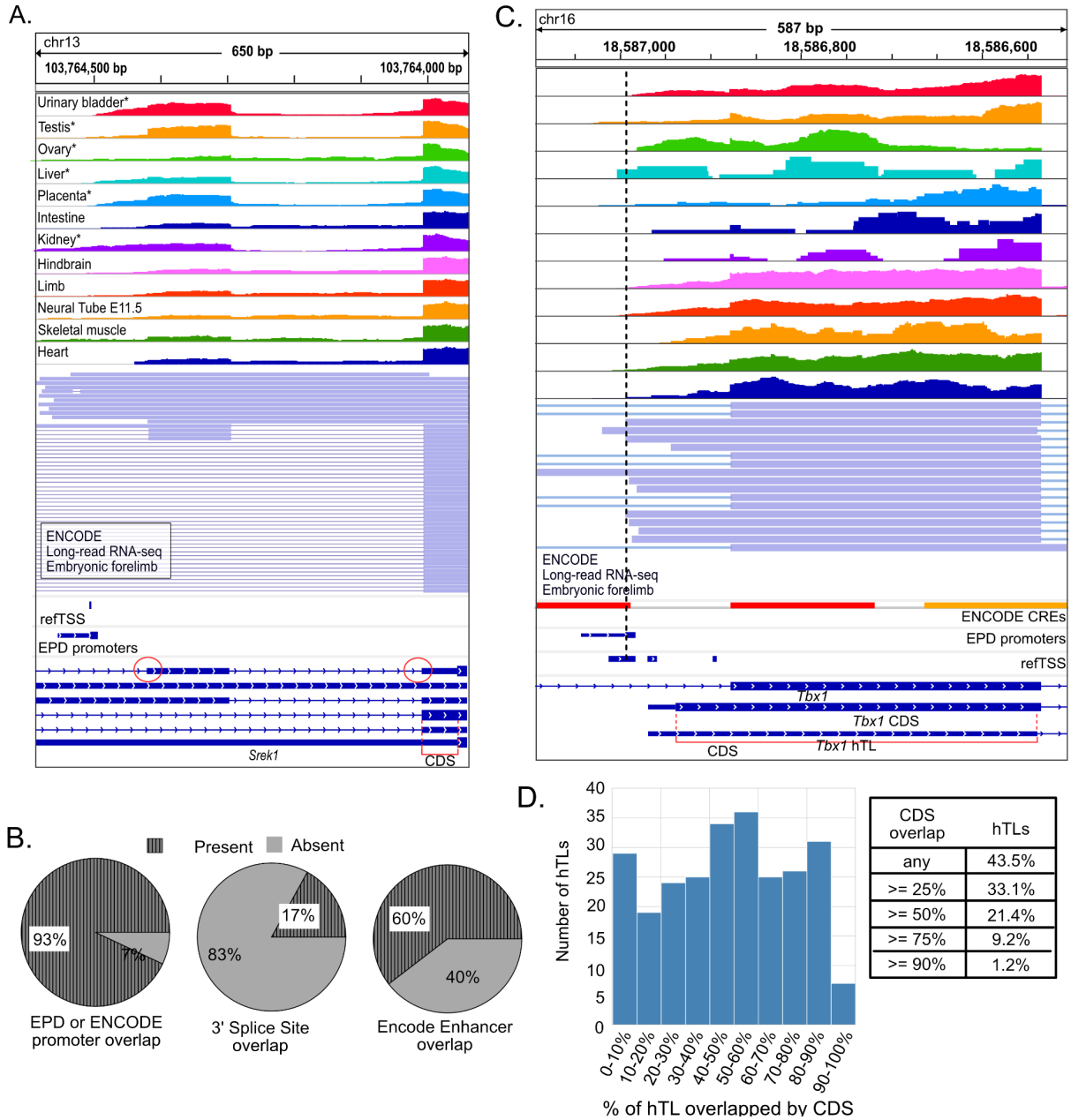


Figure S4. Annotated transcript leaders of Hoxa genes increase expression of both Rluc and Fluc. (A) Diagram of the "promoterless" bicistronic reporter plasmid pRF-ΔSV40. The putative IRES-like transcript leaders were cloned between Renilla (Rluc) and Firefly (Fluc) luciferase open reading frames and transfected into mouse mesenchymal cells. (B) Bar graphs showing raw luminescence values for Rluc and Fluc from each transfection. Error bars show 95% confidence intervals with n = 3. Most transcript leaders increase expression of both Rluc and Fluc, but to differing extents. For example, *M. mus. Hoxa9* UTR induces Fluc, but not Rluc expression well above background. The shorter "IRES-like" region induces expression of both, but Fluc is expressed more than Rluc compared to empty vector. The active (IRES-like) UTRs induce Fluc more than they induce Rluc, leading to a higher ratio (see figure 2), while the inactive UTRs induce similar fold changes in Rluc and Fluc expression, which the exception of *Hoxa13*, which induces neither gene. Note that this interpretation does not account for potential variation in active plasmid concentrations and / or transfection efficiency. However, it seems unlikely that such variation could explain the wide-ranging variation in Rluc expression.



CDS overlap	hTLs
any	43.5%
>= 25%	33.1%
>= 50%	21.4%
>= 75%	9.2%
>= 90%	1.2%

Figure S5. hTLs defined by Byeon et al. 2021 overlap promoters, splice sites, enhancers, and protein coding sequences. (A) Genome browser screenshot showing an example of a hTL from *Srek1*, which drives expression in the bicistronic reporter assay. Short- and Long-read RNA-seq data suggest transcription initiates internally in this annotated transc-ript leader. The hTL overlaps an EPD promoter, two 3' splice sites, and protein coding sequence from an alternatively spliced isoform (CDS) of the gene. (B) Pie graphs showing the percentage of hTLs that overlap EPD or ENCODE promoters, annotated 3' splices sites, and ENCODE annotated transcriptional enhancer regions. (C) Genome browser screenshot showing a hTL from *Tbx1*, which almost entirely overlaps protein coding sequence from two other annotated transcript isoforms. The annotated protein coding sequence is translated in ribosome profiling data (GWIPs-VIZ; not shown) and has PhyloP conservation scores consistent with its translation (lower scores at wobble nucleotides). (D) Histogram showing the number of hTLs with varying percentages of CDS. 256 hTLs (43.5% of all hTLs) overlap annotated CDS regions, and a third of hTLs have at least 25% of their sequence overlapped with annotated CDS.

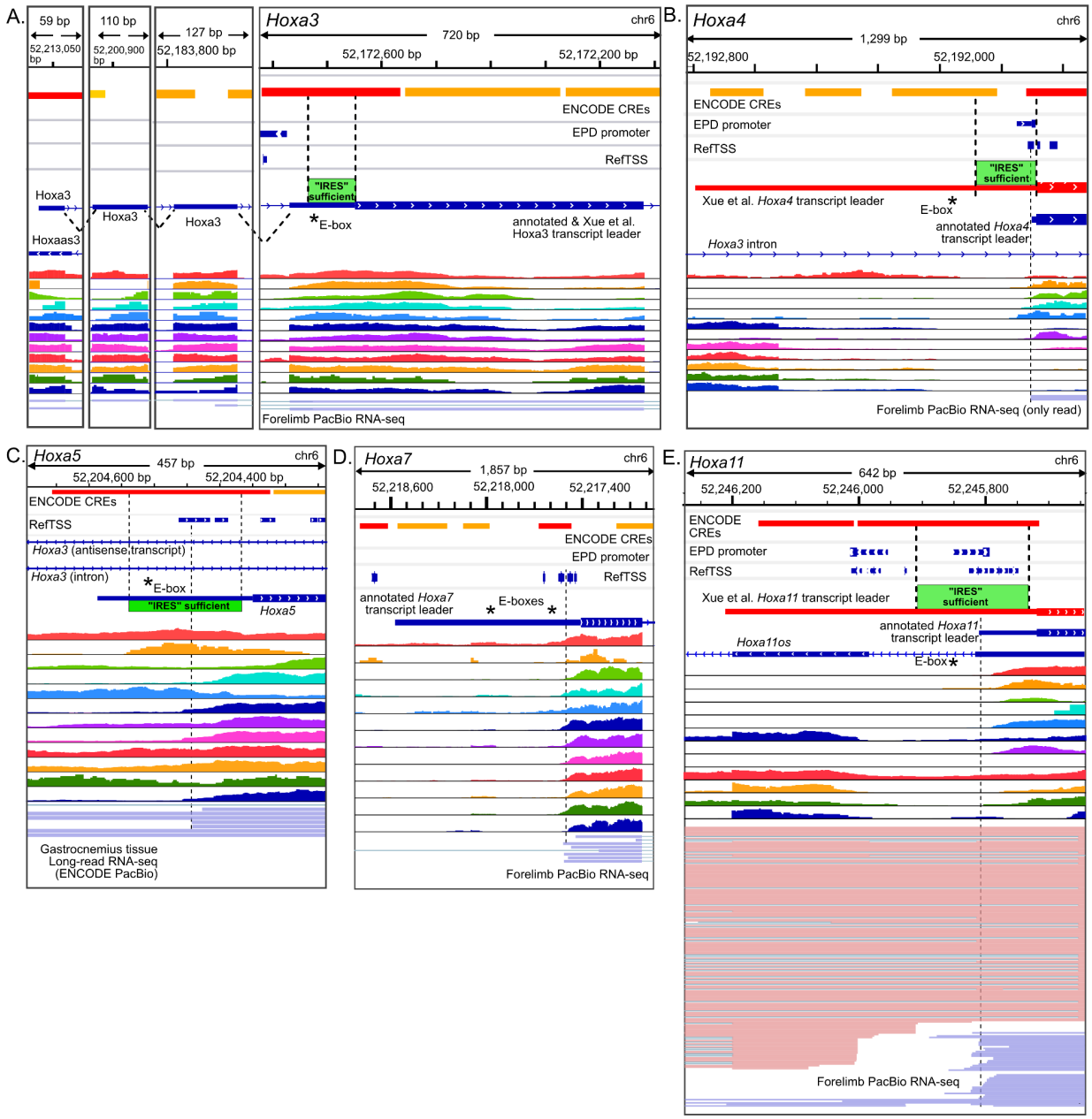


Figure S6. Misannotated transcript leaders and promoter overlap with putative IRESes from *Hoxa* genes. (A) The sufficient region of the putative *Hoxa3* IRES overlaps an ENCODE promoter. (B) The *Hoxa4* transcript leader mapped by Xue et al. 2015 is much longer than the annotated leader. The annotated, short leader is supported by short- and long-read RNA-seq data from ENCODE, and the sufficient region of the putative IRES overlaps an EPD promoter and refTSS sites. (C) The sufficient region of the putative *Hoxa5* IRES overlaps an ENCODE promoter and refTSS, which are supported by short- and long-read RNA-seq data. (D) The transcript leader of *Hoxa7* is much shorter than annotated, such that the region reported to be a putative IRES (Byeon et al., 2021) encompasses an ENCODE promoter and refTSS sites. (E) The *Hoxa11* transcript leader mapped by Xue et al. (2015) is much longer than the annotated transcript leader. The sufficient region of the putative IRES overlaps ENCODE and EPD promoters, and refTSS sites. Short- and long-read RNA-seq data support internal transcription initiation at the shorter annotated promoter. Short-read RNA-seq data are color-coded as in Figure 2. As with *Hoxa9* (Figure 2) extended, misannotated transcript leaders overlap introns in *Hoxa4*, and *Hoxa5*. * Asterisks show the locations of E-box motifs mutated in figure 3.

Figure S7. The transcript leader for mouse Beta actin is misannotated and includes a promoter. The refGene annotated transcript leader is shown. The annotated transcript leader begins with a TATA box, overlaps ENCODE and EPD promoters, and a refTSS site. Long-read RNA-seq data from ENCODE supports internal transcription initiation at the annotated promoter and refTSS site.

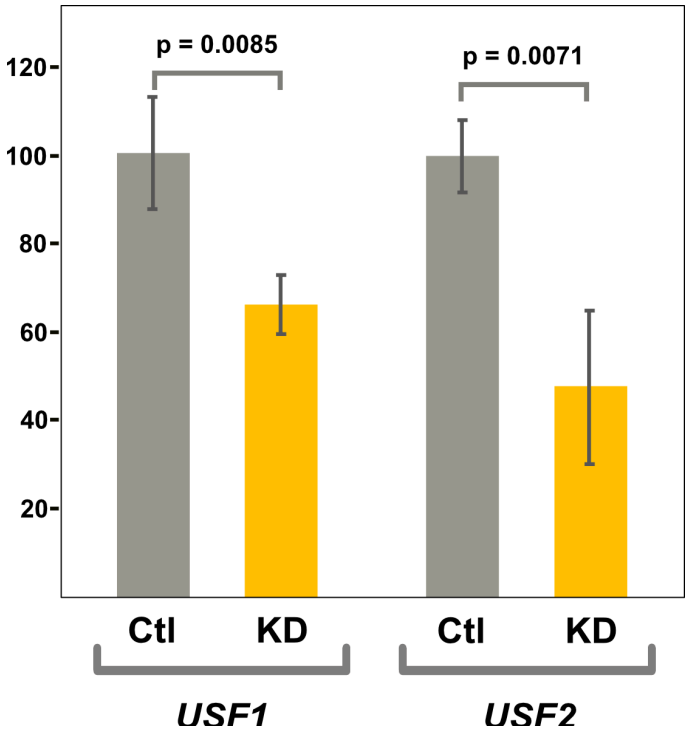


Figure S8. RT-qPCR validation of USF1/2 codepletion by siRNA. USF1 and USF2 were assayed using RT-qPCR to compare their mRNA levels from cells treated with control (scrambled) siRNA and cells treated with a mixture of USF1 and USF2 siRNA (Santa Cruz Biotechnology. Both USF1 and USF2 mRNA levels were significantly depleted, compared to the scrambled control sample. Bar graph shows the average of 3 replicates, with error bars representing 95% confidence intervals. P-values shown are from a 1-tailed Welch's T-test