The identification and functional annotation of RNA structures conserved in vertebrates

Seemann, Ernst Stefan; Mirza, Aashiq Hussain; Hansen, Claus; Bang-Berthelsen, Claus Heiner; Garde, Christian; Christensen-Dalsgaard, Mikkel ; Torarinsson, Elfar; Workman, Christopher; Pociot, Flemming; Nielsen, Henrik

Total number of authors:
13

Published in:
Genome Research

Link to article, DOI:
10.1101/gr.208652.116

Publication date:
2017

Document Version
Peer reviewed version

The identification and functional annotation of RNA structures conserved in vertebrates

Stefan E Seemann[1,2], Aashiq H Mirza[1,3*], Claus Hansen[1,6*], Claus H Bang-Berthelsen[1,4*¶], Christian Garde[1,5*], Mikkel Christensen-Dalsgaard[1,6], Elfar Torarinsson[1], Zizhen Yao[9], Christopher T Workman[1,5], Flemming Pociot[1,3], Henrik Nielsen[1,6], Niels Tommerup[1,6], Walter L Ruzzo[1,7,8], Jan Gorodkin[1,2†]

1 Center for non-coding RNA in Technology and Health (RTH), University of Copenhagen, Denmark

2 Department of Veterinary and Animal Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark

3 Copenhagen Diabetes Research Center (CPH-DIRECT), Herlev University Hospital, Denmark

4 Department of Obesity Biology and Department of Molecular Genetics, Novo Nordisk A/S, Denmark

5 Department of Biotechnology and Biomedicine, Technical University of Denmark, Denmark

6 Department of Cellular and Molecular Medicine (ICMM), Faculty of Health and Medical Sciences, University of Copenhagen, Denmark

7 School of Computer Science and Engineering and Department of Genome Sciences, University of Washington, Seattle, USA

8 Fred Hutchinson Cancer Research Center, Seattle, USA

9 Allen Institute for Brain Science, Seattle, USA

¶ Current affiliation: Microbial Biotechnology and Biorefining, National Food Institute, Technical University of Denmark, Denmark

∗ These authors contributed equally to this work.

† To whom correspondence should be addressed (gorodkin@rth.dk).

Structured elements of RNA molecules are essential in, e.g., RNA stabilization, localization and protein interaction, and their conservation across species suggests a common functional role. We computationally screened vertebrate genomes for Conserved RNA Structures (CRSs), leveraging structure-based, rather than sequence-based, alignments. After careful correction for sequence identity and GC content, we predict ~516k human genomic regions containing CRSs. We find that a substantial fraction of human-mouse CRS regions (i) co-localize consistently with binding sites of the same RNA binding proteins (RBPs) or (ii) are transcribed in corresponding tissues. Additionally, a CaptureSeq experiment revealed expression of many of our CRS regions in human fetal brain, including 662 novel ones. For selected human and mouse candidate pairs, qRT-PCR and *in vitro* RNA structure probing supported both shared expression and shared structure despite low abundance and low sequence identity. About 30k CRS regions are located near coding or long non-coding RNA genes or within enhancers. Structured (CRS overlapping) enhancer RNAs and extended 3' ends have significantly increased expression levels over their non-structured counterparts. Our findings of transcribed uncharacterized regulatory regions that contain CRSs support their RNA-mediated functionality.

Computational analyses have suggested many conserved structured RNAs in vertebrate genomes (Washietl et al. 2005; Torarinsson et al. 2008; Parker et al. 2011; Smith et al. 2013). Recent transcriptome-wide experiments also support a diverse RNA structure landscape (Ding et al. 2014; Rouskin et al. 2014; Wan et al. 2014; Aw et al. 2016; Lu et al. 2016; Sharma et al. 2016). These experiments, however, do not broadly exploit the phylogenetic context in which functionally important RNAs appear, especially compensatory base pair changes. Furthermore, previous computational screens for conserved RNA structures have focused on sequence-based alignments (Gorodkin et al. 2010), although structural alignments more sensitively capture evolutionarily conserved RNA structures (Wang et al. 2007). While structure is known to be critical to the biogenesis or function of many non-coding RNAs (ncRNAs), it remains unclear how ubiquitous a role conserved structures play. For example, a recent experiment mapping RNA duplexes in living human and mouse cells (Lu et al. 2016) reported conserved structured RNA domains in several long ncRNAs (lncRNAs), including *XIST* and *MALAT1*, while *in silico* studies based on single sequence folding and sequence-based alignments (Managadze et al. 2011) have indicated that RNA secondary structures are depleted in lncRNAs (Ulitsky and Bartel 2013). The low sequence conservation of most lncRNAs, while complicating identification of conserved RNA structures, does not preclude their existence, such as in telomerase RNA (structurally similar across vertebrates despite human-mouse sequence identity of ~60%).

Given these difficulties in detecting conserved structures, accuracy of computational screening methods are a prime concern. Here, we present a carefully designed discovery pipeline and a significantly improved scoring scheme, with careful control for technical factors such as di-nucleotide composition and GC-content, with the goal of reducing the False Discovery Rate (FDR) of the predicted Conserved RNA Structures (CRSs).

Regulatory features of RNA structures have been extensively studied in bacteria (Waters and Storz 2009) but the vast landscape of RNA regulatory elements in vertebrates remains largely uncharacterized.

Conservation, structural or otherwise, typically implies function, but does not tell us *what* function. RNA structures are known to be involved in gene regulation through transcript stabilization (Goodarzi et al. 2012), interaction with RNA binding proteins (RBPs) (Ray et al. 2013), and other processes. While most RBPs contain a few RNA-binding domains, the contextual features that regulate RBP binding are often of limited sequence specificity and are not well known. Some RBPs specifically bind double-stranded RNAs (dsRBPs); examples include DGCR8 and Dicer, important in siRNA and microRNA biogenesis (Macias et al. 2012; Rybak-Wolf et al. 2014), and STAU1 in regulated RNA decay (Kim et al. 2005). Other RBPs bind unpaired nucleotides exposed in loops and additional secondary structure elements (Lunde et al. 2007).

The complexity of the vertebrate transcriptome had been underestimated for decades, but the advent of high-throughput sequencing has enabled the identification of many new transcripts. For example, expression upstream of promoters (Seila et al. 2008; Preker et al. 2011) and RNAs transcribed from enhancers (eRNAs) (Andersson et al. 2014a; Arner et al. 2015), have recently been recognized to be common, but are often viewed as transcriptional by-products at accessible genomic sites, especially because of generally rapid degradation by the nuclear exosome (Jensen et al. 2013; Ntini et al. 2013; Almada et al. 2013). However, experimental data increasingly suggests functional roles for these transcripts (Di Giammartino et al. 2011; Rinn and Chang 2012; Li et al. 2013). As another example, the majority of human genes are alternatively cleaved and polyadenylated, and these alternative isoforms differ in their stability, localization and translational efficiency (Elkon et al. 2013). Many regulatory elements are located in untranslated regions of mRNAs (UTRs) and recognized by RBPs (Berkovits and Mayr 2015). Many are structured, as indicated by the large repertoire of RNA structure families in Rfam (Nawrocki et al. 2015). The structures themselves might even affect alternative polyadenylation and stability (Di Giammartino et al. 2011), or co-exist in downstream independent transcripts since RNA polymerase II does not cease transcription at the poly(A) site (Glover-Cutter et al. 2008).

In short, many noncoding genomic regions, including gene regulatory regions, are transcribed and may host functionally important structures, but their superficial lack of sequence conservation might systematically bias against discovery of RNA structures, thus motivating our main aim: genome-wide exploration of conserved RNA structures, based on structure-aware multi-species alignments. To assess our *in silico* CRS predictions, we present extensive correlations with public and novel experimental data in two directions. The first assesses the accuracy of our computational predictions. For example, we used RNA structure probing experiments to test our predictions of structure conservation between human and mouse, and used RNA-seq and RT-qPCR to test human-mouse co-expression. A second direction assesses potential CRS roles in specific functions, for example, RBP interactions and enhancer activity.

Results

Identification of conserved RNA structures by local structural alignments

To identify Conserved RNA Structures (CRSs), we extracted sequences from MULTIZ alignments (MAs) from 17 vertebrates (hg18) (Blanchette et al. 2004), collectively corresponding to ~50% of the human genome (Methods). RNA structure predictions were made in these putatively orthologous sequence sets by CMfinder, which locally and structurally aligns a set of unaligned sequences, discarding apparently irrelevant ones (Yao et al. 2006). CMfinder is not constrained by the initial sequence-based alignment or by pre-defined window sizes. It has been broadly successful, e.g., aiding the discovery of large bacterial ncRNAs (Weinberg et al. 2009), and of numerous ribozymes and riboswitches (Weinberg et al. 2007). Predictions from the 17-species analysis were extended to the 100-species tree (hg38) (Methods).

We predicted 773,850 CRSs (pscore≥50, Methods) covering 515,506 CRS regions (genomic regions of overlapping CRSs). We estimated our CRS False Discovery Rate (FDR) to be 14.1±5.1% within the most common GC-content range of 20% to 65% (using a di-nucleotide controlled and GC-content corrected

phylogenetic null model; Supplemental Methods), while the top 20% of CRSs ranked by pscore have an estimated FDR<10% (Fig. 1A, Supplemental Fig. S1). A GC-content-specific FDR is important because GC contents vary strongly among CRSs and across different biotypes (genomic locations of similar characteristics) (Fig. 1B).

72% of CMfinder-predicted base pairs agree with an independent *in vivo* biological assessment by genome-wide structure probing (Rouskin et al. 2014) (Methods). The even higher agreement between our *in silico* predictions and their *in vitro* data (Supplemental Fig. S2) further supports our methodology; neither *in silico* nor *in vitro* consider the cellular environment, including protein binding.

Predicted CRSs average 71±46bp in length and cover 36.5 million bases (~2.6%) of the human input sequence. On average, they are conserved in 45±19 species of the 100-species tree (Fig. 1F,I) with deeper conservation in sncRNAs and mRNAs (coding sequences (CDS)). CRSs regions are mostly intronic or intergenic (Fig. 1D; annotation sources are listed in Supplemental Material), and are enriched for small ncRNAs (sncRNAs; including 230 precursor-microRNAs and 199 snoRNAs) and UTRs (Fig.1C, Supplemental Table S1). They overlap 36% of the 1,067 structured (base pair content>30%) Rfam (Burge et al. 2013) input sequences, comprising mostly sncRNAs and cis-regulatory structures in UTRs. The majority of lncRNAs lack CRSs (Fig. 1C; Supplemental Fig. S3) consistent with previous observations (Ulitsky and Bartel 2013). Nonetheless, in addition to known examples such as tRNA-like structures in *MALAT1* and *NEAT1* (Zhang et al. 2014) (Supplemental Fig. S4), many lncRNAs host CRSs, including 22% of screened lncRNAs annotated in GENCODE v25 (Harrow et al. 2012), 19% of lncRNAs annotated from RNA-seq data (PLAR) (Hezroni et al. 2015), 31% of antisense ncRNAs, and 30% of processed pseudogenes. Within lncRNAs, CRS density decreases from 5' to 3' (Fig. 1E). A small number of CRSs (13,535) outside annotated coding sequences (GENCODE) hold coding potential according to PhyloCSF (Lin et al. 2011). Although 167k CRSs (21.6%) overlap repeats flagged by RepeatMasker v4.0.5 or TandemRepeatFinder v4.0.4 (Smit et al. 2013), almost all repeat families are depleted of CRSs

(Supplemental Table S2). SINE, LINE and simple repeats comprise the majority of CRS/repeat overlaps, including 1,572 CRSs that overlap Alu elements.

Evolutionary constraints on RNA structure do not necessarily coincide with evolutionary constraints on sequence. In particular, ~50% of CRSs fall outside of conserved elements identified by phastCons (Siepel et al. 2005) in the 100-species alignment, and CRS regions of low Sequence Identity (SI) showed a higher degree of re-alignment when structure was taken into account (Fig. 1G,H). CRS regions within annotated coding sequences, UTRs or targets of RNA- and DNA-binding proteins (e.g., transcription factor (TF) binding sites) generally show higher SI and less re-alignment. In contrast, intronic CRS regions and ones within 2kb upstream ("5' extension") or within 2kb downstream ("3' extension") to mRNAs and lncRNAs, showed lower SI and significantly more re-alignment ($P < 10^{-6}$, *t*-test; Fig. 1G,H). Expression extending beyond annotated UTRs and lncRNAs was repeatedly observed in transcriptomic data, which is addressed below. Lower SI in gene extensions (~64% SI) may indicate faster adaptation of these RNA structures to novel functions than those in either lncRNAs (~66% SI) or mRNAs (CDSs and UTRs ~70% SI).

Purifying selection

Despite their somewhat low SI, CRSs show signatures of purifying selection. Firstly, nucleotide distances between primates and rodents are lower in CRSs than in nearby ancestral repeats (ARs) or intergenic loci ($P < 10^{-12}$, two-sided Kolmogorov-Smirnov test; Supplemental Fig. S5A,B). Nucleotide substitution patterns for ~62% of the 26k CRSs with orthologs and appropriate nearby control sequences are improbable under neutralist models (Methods). Secondly, CRSs are enriched within DNA that has been subject to purifying selection with respect to indels (Lunter et al. 2006) ($P < 10^{-10}$, one-sided *Z*-test, Benjamini-Hochberg (BH) corrected; Supplemental Fig. S5D). Thirdly, the minor allele frequencies (MAFs) in a large human population (Auton et al. 2015) of 271k CRSs ≥2kb away from known mRNAs

are significantly lower than in regions 5 to 10kb up- and downstream ($P < 10^{-16}$, Mann-Whitney *U* test, Supplemental Methods).

Co-localization between CRSs and conserved RBP binding sites

RNA targets from *in vivo* CLIP experiments for 67 human RBPs overlap 102k CRS regions (Methods), which correlates with CRS enrichment for binding sites of 76% of human RBPs after stratifying for GC content and SI ($P < 10^{-10}$, one-sided *Z*-test, BH-corrected). For most RBPs, CRSs were enriched around binding sites in regions having GC content >40% or high SI (Supplemental Fig. S6). For some RBPs, CRSs were also enriched in low SI regions. Examples include the IGF2 mRNA binding protein family (IGF2BPs), whose three mammalian paralogs contain two RRM and four KH domains, all putatively involved in RNA binding with limited sequence specificity. Human-mouse conserved CRSs associated to IGF2BP2 binding in human also showed RNA binding in mouse, supporting their functional roles in binding site recognition. More generally, 7 of 10 studied RBP orthologs in mouse (Methods), including IGF2BP2, are enriched for binding sites overlapped by CRSs both in human and mouse ($P < 10^{-7}$, Fisher's Exact Test (FET), Fig. 2A). Alternative splicing is an example of RBP recruitment modulated by the kinetics and thermodynamics of RNA structure (Raker et al. 2009) and in our study the binding sites for the splicing factors FOX-2, HNRNPA1 and PTB were enriched for CRSs near splice sites ($P < 0.001$, FET).

Conserved expression in human and mouse

Cross-species conserved transcriptional activity of CRSs can imply conserved biological function. We selected closely matched human/mouse RNA-seq samples from 10 tissues (Supplemental Methods, Supplemental Table S8). In both species, the highest expression levels of CRSs occurred within exons of mRNAs and lncRNAs (Supplemental Fig. S7). Using an empirical *P*-value calculated from background expression (Methods), conserved transcriptional activity ($P < 0.01$) was supported for ~36% of the

shared human-mouse CRSs having concordant biotypes in both species (Fig. 2B). This was dominated by CRSs of mRNAs and introns (~50%), but CRSs within enhancers, 5' extensions, 3' extensions and lncRNAs were also well-represented (>20% of shared CRSs of each biotype were co-expressed). This overlap in expression remains evident in individual tissues. E.g., among CRSs expressed in at least 2 tissues in both species, 23% showed strongly correlated expression (Pearson's correlation $r \geq 0.8$), including 164 from lncRNAs, 788 from enhancers, and 780 intergenic CRSs (Fig. 2C and Supplemental Fig. S7G, Methods). Despite relatively small numbers, CRSs within mRNAs, lncRNAs, sncRNAs, 5' and 3' extensions have significantly larger cross-species co-expression than background ($P < 0.05$, one-sided Mann-Whitney *U* test, BH-corrected). For example, the lncRNA *MIR22HG*, hosting several CRS regions in addition to the microRNA *MIR22*, is expressed in all tissues considered here, and the noncoding testis development-related gene 1 (*TDRG1*), exclusively expressed in testis of both human and mouse, has a large CRS region of low SI (~60%) extending beyond its annotated 3' end.

A CaptureSeq experiment detects weakly expressed structured RNAs

Almost 50% of the CRS regions overlap with abundant transcription in publicly available total RNA-seq and poly(A) RNA-seq from 16 and 19 human tissues, respectively (empirical *P*-value < 0.01 and normalized count per million reads (CPM/RLE) > 1 in ≥ 2 tissues; Methods). In standard RNA-seq experiments, however, weakly expressed transcripts are undetectable or indistinguishable from nonspecific transcription (Lin et al. 2014). To improve detection sensitivity, we designed capture probes for 77,320 CRS regions (Methods, Supplemental Methods) and coupled the capture with an RNA-seq experiment (CaptureSeq). In Fig. 3A (and Supplemental Fig. S8) we observe good agreement between the expression of the captured CRS regions and publicly available data. In human fetal brain alone 8,385 CRS regions were significantly expressed ($P < 0.1$; Methods, Supplemental Methods) including 662 transcripts that were previously not detected in brain (<1 CPM/RLE; Methods). The majority of these CRS regions are located in UTRs or UTR extensions, another 115 CRS regions fall in intergenic regions,

205 in lncRNAs and 50 in sncRNAs. Many of these regions (1,475) are weakly conserved in sequence (SI<60%). Examples include human-mouse conserved RNA structures that overlap the known conserved stem loop in exon 4 of *XIST* (Fang et al. 2015), the highly structured telomerase RNA element, the brain-specific *mir9-2* host gene *LINC00461* and the *SH3RF3* antisense RNA 1 (*SH3RF3-AS1*).

qRT-PCR shows correlated expression profiles in human and mouse

To further explore conserved expression between human and mouse presented above, we compared expression of selected CRSs across 7 tissues in both species via qRT-PCR. We studied 23 CRS regions of low to medium SI and with weak expression in publicly available brain samples (Methods) covering 5 novel transcripts, 5 transcripts close to annotated mRNAs, 8 recently annotated extensions of UTRs and 5 lncRNAs (Supplemental Table S3). These regions were detected in at least 80% of the examined tissues despite low expression levels, and 9 CRS regions showed strong co-expression between human and mouse (Pearson's correlation $r > 0.8$, Fig. 3B). For example, Fig. 3C shows a structure conserved in 59 species with SI of 75% that was predicted in the tissue-specific lncRNA *AC073046.25*. Thus, our CaptureSeq strategy identified (and qRT-PCR partially validated) conserved expression profiles of CRSs with low abundance and low sequence conservation, leading to putative functional genes.

Structure probing shows conserved RNA structures in human and mouse

To investigate whether CRSs of low SI were indeed structurally conserved, we performed RNA structure probing (Methods) of homologous human and mouse sequences of 10 CRSs, selected based on their low FDR (~10%), low SI and qRT-PCR-validated co-expression in human and mouse brain. *In vitro* transcription yielded four CRS pairs suitable for structure probing as determined by native gel electrophoresis (Supplemental Table S4). In all four cases there was strong consistency between the predicted structures and the experimental analyses (Fig. 4, Supplemental Fig. S9). The CRSs originated from the 3' end of the brain-specific mRNA *KCNG2*, the short 5' UTR of brain-specific *EOMES*,

lncRNA *MIR4697HG* (the hosted microRNA was not probed), and CRS candidate M1695693, found downstream of the annotated 3' end of *HOMER2* (a postsynaptic density scaffolding protein). Poly(A) signals from RNA-seq (Gruber et al. 2016) supported an extended 3' UTR of *HOMER2* covering the CRS region (Fig. 4A). Despite 45% SI between human and mouse, the probing showed that the two dissimilar sequences can fold into closely related structures largely in agreement with the structural alignment (Fig. 4B,C).

RNA structures are enriched within gene regulatory regions

A substantial fraction (~40%) of the 433k intergenic and intronic CRS regions are located at TF binding sites, DNase hypersensitive sites (DHSs) or loci exhibiting promoter- and enhancer-specific chromatin marks, all suggesting regulatory activities. This prompted a more detailed analysis of the ~30k CRS regions found within (1) 1kb of enhancers, (2) 5' extensions or (3) 3' extensions of mRNAs and lncRNAs, collectively called gene regulatory regions here. We attempted to control for several potential confounding factors in these regions. E.g., their comparatively rapid evolution (lower SI than other CRSs ($P < 10^{-7}$, two-sided Mann-Whitney *U* test), Fig. 1G) complicates the use of phylogenetic measures (Ponting 2008) and the significantly higher GC content compared to other CRSs ($P < 10^{-78}$, two-sided Mann-Whitney *U* test) makes GC content correction especially important. Since approximately half of these CRSs co-localize with TF binding sites, we considered, but ruled out, the possibility that palindromic DNA sequences of TF binding sites might give rise to spurious CRS predictions (Supplemental Fig. S10). Our study nonetheless reveals not only structure conservation in these gene regulatory regions, but shows CRSs to be sharply enriched (Fig. 5A,E,I; Supplemental Table S5), even after allowing for a slightly higher GC-content-corrected FDR for predictions around TSSs (Fig. 5E, lower subpanel).

To further preclude false signal from DNA motifs unrelated to RNA structure, we analyzed in greater detail the subset of CRSs located near experimentally determined but unannotated transcript boundaries (TSSs and poly(A) sites). Using DHSs, CAGE measured TSSs and characteristic chromatin marks (Methods; Supplemental Fig. S11), we found 10,110 enhancer transcripts of which 2,862 contained CRSs (with >50% of their length downstream of the TSS) (Fold Enrichment FE=1.67, *Z*-test $P < 10^{-185}$; Fig. 5A). Fig. 6A shows two putative eRNAs in a structured intergenic region of low SI between two protein-coding genes involved in glucose metabolism. We also found 1,077 mRNAs and 129 lncRNAs with transcripts upstream and antisense to their annotated TSS, of which 337 contained CRSs (FE=1.83, *Z*-test $P < 10^{-40}$; Fig. 5E). However, upstream sense TSSs (≤1,300bp from annotated TSSs; Supplemental Fig. S11) for a considerably larger number of mRNAs (2,530) and lncRNAs (519) plausibly reflect alternative TSSs (unannotated in GENCODE), and 24% of these 5' extensions contained CRSs. In four instances the CRS region overlapped pre-miRNAs (*MIR320A*, *MIR34B*, *MIR219A1*, *MIR4665*), suggesting exploitation of TSS upstream transcripts (of either orientation) to co-regulate multiple components of specific pathways. For example, *MIR320A* is transcribed antisense to and located in the promoter region of *POLR3D* and regulates transcription factors of *POLR3D* (Kim et al. 2008) (Fig. 6B). Fig. 6C shows a bidirectionally transcribed locus of unknown function at the promoter of an lncRNAs. Active poly(A) site data (Gruber et al. 2016), revealed 2,885 mRNAs and 1,260 lncRNAs with an alternative poly(A) site between 50bp and 2kb downstream of the most distal GENCODE annotated 3' end. Of these, 543 mRNAs and 203 lncRNAs had a predicted CRS within their 3' extension, reflecting a modest FE of CRSs in the mRNA extensions (FE=1.15, *Z*-test *P*=0.0001; Fig. 5I). An example is CRS M1695693, which is likely linked to *HOMER2* as discussed above and in Fig. 4. In all three regulatory regions we saw higher density of CRSs in loci supported by experimentally defined transcript boundaries (Fig. 5A,E,I, "transcribed" curve).

CRSs impact transcription of enhancers and 3' extensions

We then explored whether the observed enrichment of predicted CRSs in regulatory regions can be linked to transcription, focusing on regions with experimentally defined transcript boundaries. Note, however, that regulatory programs are highly tissue specific, emphasizing that our conservative candidate list may only represent a subset of the transcriptomic landscape. Specifically, we examined total RNA-seq and poly(A) RNA-seq of 19 and 16 tissues, respectively (Methods). For (1) transcription around enhancers, (2) TSS-upstream transcription, and (3) transcription of 3' extensions, CRS-overlapping ("structured") transcripts show significantly higher expression levels in all tested tissues compared unstructured ones ($P < 0.001$, one-sided Mann-Whitney *U* test, BH-corrected; Fig. 5B,F,J). However, we also observed that GC content and SI were increased in structured versus unstructured transcribed regions (Fig. 5C,G,K and Supplemental Fig. S12). To account for this potential confounder we ran enrichment tests for different ranges of GC content and phastCons score and observed the following ($P < 0.05$, one-sided Mann-Whitney *U* test, BH-corrected; Supplemental Table S6): Expression of eRNAs and 3' extensions overlapping CRSs with GC content between 25% and 75% is significantly increased in most tissues, whereas TSS-upstream (sense and antisense) transcription is rarely linked to CRS overlap. The data shows that conserved structure predictions add significant information for distinguishing transcriptionally active regulatory sites from silent ones despite the impact of GC content and sequence conservation.

The positive correlation of CRSs and expression might be due to enhanced transcription and/or increased stability (slower degradation). To disambiguate these alternatives, we examined CAGE data for transcription initiation at DHSs in control vs exosome-depleted HeLa cells (Andersson et al. 2014b). Defining "stability" based on the change of expression level after exosome depletion (Methods), we find that stable eRNAs of preferentially unidirectional transcription were enriched for transcripts containing CRS regions (P=0.002, FET, BH-corrected; Fig. 5D). (We cannot rule out some impact of GC content and sequence conservation since the small sample size precludes GC-content-specific enrichment tests.)

An example of a unidirectional stable eRNA is depicted in Fig. 6D. The relationship between transcript stability and CRS presence in eRNAs may point to RNA structure alone as a contributor to stability, but the association of CRSs to stable TSS-upstream transcripts was not significant (Fig. 5H), implying that structure *per se* does not confer a stability advantage. We also note that thousands of CRS regions overlap RBP binding sites in promoter and enhancer regions (Supplemental Fig. S13), raising the possibility that CRSs mediate transcript stability by protein recruitment.

Discussion

Although the RNA structure landscape is recognized as an important feature of the transcriptome, a global analysis of its functional impact in vertebrates is still missing. In our study, we present a comprehensive screen for conserved RNA structures based on local structural alignments of human and other vertebrate genomes. In general, and in agreement with observations in (Torarinsson et al. 2008), our approach substantially re-aligned many sets of orthologous sequences, exposing well-defined structures with lower sequence identity than the input, but phylogenetically broader scope than was visible using earlier purely sequence-based approaches. Of the 774k CRSs covering 516k contiguous genomic regions, 276k showed an average pairwise sequence identity below 60% over the 17 representative genomes in the phylogenetic tree. The CRSs were enriched in a spectrum of known functional elements, supporting their global functional importance. Clustering conserved RNA structures may help identify and dissect common RNA structures shared across multiple classes of RNAs (Parker et al. 2011; Miladi et al. 2017) that may have common functionality, such as specific protein-RNA binding sites.

Most CRSs may be weakly or narrowly expressed, hence not easily detectable by ordinary transcriptome-wide RNA-seq. We designed a custom capture chip to study them further and found 8,385 CRS regions expressed in human fetal brain of which 662 were not in the publicly available brain data sets that we

examined. The qRT-PCR and *in vitro* structure probing in human and mouse of a small but diverse subset of CRS regions revealed tissue-specific expression profiles and conserved secondary structures. These results support both our *in silico* predictions and CaptureSeq strategy for *de novo* discovery of structured RNA. Overall, our CaptureSeq detected expression for >10% of the ~77k probed CRS regions, of which ~1% were novel expressed CRSs in just a single tissue, human fetal brain, demonstrating the concept. Our CaptureSeq strategy aims to detect novel RNAs on a transcriptome-wide scale. Similar approaches which tile entire genes (Mercer et al. 2014) are nicely complementary, as subsequent tiling around newly discovered expressed CRSs would meaningfully extend our knowledge about these transcripts.

We found CRS regions to significantly overlap functional elements, such as binding sites for RBPs (102k), eRNAs (15k), extended 3' ends (8k), and extended 5' ends (8k). Widespread enhancer- and TSS-upstream-antisense transcription are a still-recent observation from high-throughput sequencing; whether these transcripts have functional roles or are just noise remains controversial. Sometimes these transcripts produce lncRNAs that may contribute to the regulatory function of the genomic site (Rinn and Chang 2012). One crucial role of bi-directionally transcribed promoters and enhancers may be in transcript stabilization (Goodarzi et al. 2012). This process is relatively independent from the primary sequence, and instead may be linked to RNA structure, e.g., through protein bindings. Based on our analyses we conclude that (1) CRSs are associated with increased stability in certain genomic contexts, which at least partially explains higher abundance of structured elements seen in RNA-seq, but (2) the more general observation of greater abundance of structured elements (for a large range of GC contents) suggests that CRSs have functional roles above and beyond modulating stability.

In further support for functional roles of the CRSs, we observed that a substantial fraction of CRSs (~36% of the 433k tested) were co-expressed in human and mouse, including CRSs in transcribed regulatory regions. Although the fraction of CRSs co-expressed in corresponding human and mouse tissues was low, it was significant for CRSs of several biotypes. Furthermore, it is likely that we underestimated this

overlap because of low expression levels, differences in the biological material (tissues) and cross-platform differences between the experiments, which complicate all cross-species expression analyses.

The presence of structured RNAs of relatively low sequence conservation in lncRNAs (Fig. 1F) agrees with the observation that lncRNAs appeared recently in evolution (Rands et al. 2014; Washietl et al. 2014). A similar search, e.g., by FOLDALIGN (Sundfeld et al. 2016), in primates or other closely related species will likely elucidate more novel CRSs in lncRNA. Lower CRS sequence identities than in lncRNAs were observed in intergenic regions with signatures specific for active regulatory elements, supporting the idea that such structures play a role in ongoing evolution of transcriptional regulation.

Our computational screen complements large scale experimental efforts to probe for RNA structures (Rouskin et al. 2014; Wan et al. 2014). These experimental approaches are limited to elucidating the structure propensity of single nucleotides, and do not provide evidence for the base interaction map. Base-interactions in human and mouse could potentially complement existing sequence alignments to build a base-interactome map (Lu et al. 2016). The substantial fraction of long-range interaction from such experiments could complement the short-range interactions from CRSs and thus together provide a more complete picture of the RNA structurome.

To conclude, our CRS screen is to our knowledge the first genome-wide screen in vertebrates explicitly based on local structural alignments that does not make rigid use of pre-generated sequence-based alignments. In combination with CaptureSeq, it has revealed RNAs not detectable by standard RNA-seq experiments, and has the potential to reveal many more when repeated in other tissues and biological conditions. The CRSs themselves show evidence for purifying selection, and co-localize with a range of known functional elements, especially in enhancers and near annotated gene boundaries. Similarly, we found CRSs overlapping numerous RBP binding sites for which RNA structures have not previously been reported. Thus, our study provides support for the existence and widespread functional importance of a

broad landscape of novel RNA structure candidates widely conserved in vertebrates. Fully elucidating their roles will entail significant follow-on work.

Methods

Genome-wide screen for CRSs

CMfinder (Yao et al. 2006) locally aligns, folds and describes predicted CRSs through covariance models using an expectation-maximization style learning procedure. To predict CRSs anchored in the human genome, hg38, we carried out the following steps: (1) filtered human-based 17-species MULTIZ alignment (MA) from UCSC (http://hgdownload.cse.ucsc.edu/goldenPath/hg18/multiz17way) for length ≥60bp blocks containing human and ≥3 non-primates, resulting in 8,131,488 MA blocks covering 46% of the human genome; (2) re-aligned and predicted shared structure in these blocks using CMfinder (version 0.2.1 and 0.3 for pscore calculation for the vertebrate tree) with default parameters (maximal base pair span of 100bp: -M 100) in both reading directions; (3) used UCSC liftOver (http://hgdownload-test.soe.ucsc.edu/goldenPath/hg38/liftOver/) across genome builds to map coordinates of high-scoring (see next section) human CRSs from hg18 to hg38; (4) used liftOver across species to map human hg38 coordinates to orthologous regions in each of the other 99 vertebrate genomes in UCSC's 100-species alignment; (5) searched each such sequence (extended 50bp up- and downstream) for hits using the CRS covariance models with Infernal cmsearch (Nawrocki and Eddy 2013); (6) aligned these hits to the CRS covariance model with Infernal cmalign. Step 2 of the screen alone took more than 150 CPU years on a linux cluster (each node with two Intel Xeon E5649 2.53GHz - Westmere -EP and 24GB memory). See Supplemental Methods for our rationale in choosing this strategy.

Scoring scheme

The probabilistic ranking statistic, pscore, extends the phylo-SCFG approach of Evofold (Pedersen et al. 2006). Like EvoFold, it contains both a single-nucleotide model (a general-time-reversible model of sequence evolution on the 4 RNA bases) and a structured RNA model (analogous, on the 16 potential base pairs). A third model, also single-nucleotide, captures poorly conserved (neutral or mis-aligned) regions. We predict structures where the structured RNA model outscores both unstructured alternatives. (Vertebrate genomes are heterogeneous mixtures of well- and poorly conserved regions. Including the third model avoids many potential false positives where poorly conserved regions score poorly, but by chance the structured model happens to outscore the unstructured-but-conserved model.) We also significantly revised EvoFold's parameterization of non-canonical base pairs, added a quasi-stationary model of base-pair indels and other gaps, and reduced the number of free parameters (22 versus EvoFold's 32). Parameters were trained on structure-annotated alignments (Wang et al. 2007). All three models are scored by maximum likelihood (Felsenstein 1981) on the 17-species vertebrate tree learned by phastCons (Siepel et al. 2005). Incorporation of folding energy is the final significant departure from EvoFold. Weighting SCFG posteriors by the thermodynamic partition function emphasizes co-varying columns in structurally stable vs unstable contexts. Overall, this approach decreases the GC content bias seen in (Torarinsson et al. 2008) and sharply reduces estimated FDR compared to that approach, to RNAz, and to EvoFold, as seen in the benchmarks reported in (Yao 2008), Chapter 4. See also Supplemental Methods.

Mapping of genome-wide structure probing

We retrieved data series GSE45803 (Rouskin et al. 2014) from the NCBI Gene Expression Omnibus (GEO), trimmed adapters and low quality 3' ends (Phred33<20) using the FASTX-toolkit (http://hannonlab.cshl.edu/fastx_toolkit/), and filtered for length ≥20 nt and average Phred33>20. Preprocessed reads were mapped to hg38 using BWA-MEM (https://arxiv.org/abs/1303.3997) disallowing 5' soft-trimming (bwa mem -L 10000,5). The number of mapped reads initiating one

nucleotide 3' of each base position of the reference was calculated for the respective termination assays of the dimethyl sulphate (DMS) reaction of native RNA and denatured RNA. Read counts were normalized to sequencing depth, and the log-fold change was calculated using a pseudo count of 5 to regularize low coverage regions: ($\log_2$(denature+5) - $\log_2$(native+5)). Positions displaying a log-fold change larger than 1 were considered to be paired nucleotides. The CRS consensus structures were evaluated at nucleotide resolution using this genome-wide structure probing as a gold standard. See Supplemental Methods.

Annotation enrichment corrected for sequence identity and GC content

Statistical significance tests of CRS (CRS region) enrichment for genomic features reflect only the part of the genome corresponding to the MULTIZ alignment input set, and reflect careful control for GC content and SI. A normal approximation to the binomial distribution (one-sided *Z*-test, BH-corrected, "pnorm" function in R (R Core Team 2016)), $N(\mu = np, \sigma^2 = np(1-p))$, was used to estimate a *P*-value based on the observed overlap count *q* (between middle coordinate of CRS and genomic feature, ignoring strand information), where *p* is the probability that a CRS overlaps a feature and *n* is the number of CRSs. The statistic was only calculated if the genome (bin) covered by the feature totaled at least 1kb. We studied the CRS enrichment binned by GC content and by SI (denominator is gap-included alignment length) (or phastCons) where each was calculated for 100bp windows of concatenated MA blocks. Enrichment tests were repeated for CRSs filtered to remove repeat and (semi-)inverted repeat sequences resulting in the same conclusions.

Evolutionary selection analysis

Selection in CRSs was tested in three ways. Firstly, we estimated pairwise base distance (Ponjavic et al. 2007) between human, mouse and macaque sequences using baseml (http://abacus.gene.ucl.ac.uk/software/paml.html) with model REV/GTR for CRSs (d_{CRS}), ARs (d_{AR}) and intergenic loci (d_{Inter}) after removing gap columns in CRS alignments and 17-species MULTIZ alignments (ARs, intergenic). Purifying selection was defined as both selection ratios d_{CRS}/d_{AR} and

d_{CRS}/d_{Inter} being smaller than 0.95 (15,131 CRSs). Secondly, CRSs enrichment inside indel-purified segments (IPSs) (Lunter et al. 2006) was conducted as described in Methods section "Annotation enrichment corrected for sequence identity and GC content". Thirdly, using whole genome sequencing data from Phase 3 of the 1000 Genome Project (Auton et al. 2015), we analyzed the MAFs of polymorphisms in CRSs (low coverage .vcf from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/). See Supplemental Methods and Supplemental Fig. S5.

CLIP-seq analysis

CLIP data for 67 human RBPs and 10 mouse orthologs were collected from public databases (Supplemental Table S7). Reads were preprocessed using cutadapt v1.2.1 (http://journal.embnet.org/index.php/embnetjournal/article/view/200) and mapped to hg38 or mm10 with Bowtie 2 (Langmead and Salzberg 2012). PCR duplicates were removed using Picard v.197 (http://broadinstitute.github.io/picard/) and peaks called ($P < 0.01$) using Piranha v1.2 (Uren et al. 2012). Enrichment studies for CRSs in human RBP binding sites were conducted as described in Methods section "Annotation enrichment corrected for sequence identity and GC content". Enrichment for CRSs in human-mouse conserved RBP binding sites was tested using FET (contingency table: CRS conservation vs RBP binding site conservation). A mouse RBP binding site falling within 50bp of the human counterpart after liftover from mm10 to hg38 was considered conserved.

Expression analysis

Premapped reads (.bam) to the human genome (hg38) were analyzed from publicly available total RNA-seq libraries of 19 tissues (ENCODE phase 3 ("The ENCODE Project Consortium" 2012)) and poly(A) RNA-seq libraries of 16 tissues (Illumina Human Body Map 2.0 (HBM)). All libraries are listed in Supplemental Table S8. Uniquely mapped reads were counted for overlap with CRS regions (201bp window around the center of CRS region) using featureCounts v1.5.0-p1 (parameters -s 0 -T 8 -Q 10 -p -

P -d 50 -D 300 -C -B --read2pos 5) (Liao et al. 2014). To define library specific cutoffs for expression, we calculated an empirical *P*-value based on the read count distribution of random genomic loci (201bp window) from MULTIZ alignment input set. Regions whose read counts have $P < 0.01$ were considered to be expressed. Read counts were converted to count per million mapped reads (CPM) and normalized between libraries using the relative log expression (RLE) normalization procedure in edgeR (Robinson et al. 2010). In case of replicates, we calculated mean values for normalized read counts for each tissue.

CaptureSeq experiment

We designed 125,000 probes, each 60bp long, covering both strands, more than three mismatches to their closest genomic paralog, representing 77,320 CRS regions with largest pscores and conservation between human and mouse. More than 70% of probed CRS regions were intronic or intergenic. Total RNA from human fetal brain (Clontech) was DNase I treated (Invitrogen), rRNA was depleted using RiboMinus (Invitrogen) according to the manufacturer's recommendations, cleaned on Microcon YM-30 columns (Millipore), chemically fragmented and cleaned on Microcon YM-10 columns (Millipore). Fragmented RNA was reverse transcribed using a random hexamer with an attached adaptor. Following reverse transcription, second strand synthesis was performed, blunt ended and adaptor ligated. The library was size-selected, 100-200bp were excised and cleaned. Excised fragments were enriched by 18 cycles of PCR and cleaned (Qiagen). The library was split into two equally sized sub-libraries for investigating two different annealing temperatures to optimize the hybridization step of potentially self-folded RNA probes. The dried library was mixed with hybridization buffer and denatured, immediately loaded onto the custom chip (NimbleGen) and incubated at 70°C for 3 days. The slide was eluted according to the manufacturer's protocol (NimbleGen Arrays User's Guide, Sequence Capture Array Delivery, Version 3.2). Following elution, the samples were enriched by 19 cycles of PCR and cleaned (Qiagen). The chip was stripped and re-hybridized with the second half of the library at 42°C for 3 days. The eluate was washed and enriched as described above.

Analysis of CaptureSeq

Reads (.fastq) were trimmed for low-quality 3' ends (Phred33<30) and adapters, and trimmed reads shorter than 40 nt were discarded using cutadapt v1.8.3. Cleaned reads were aligned to the human genome (hg38) using STAR 2.5.2a (default parameters) and alignments of ≤2 mismatches were reported (--outFilterMismatchNoverLmax 0.03). The following strategy defined read islands, their read counts and significance of RNA probe assigned read islands (on-targets): (1) filter uniquely mapped reads and remove simple repeats (≥50% overlap); (2) extend reads to 150 nt in reading direction (unified reads) and define a region of overlapping unified reads as a read island; (3) count reads inside a read island; (4) intersect read islands with RNA probes (overlap ≥1 nt); (5) calculate empirical *P*-value for read counts of RNA probe assigned read islands. The empirical *P*-value is based on the read count distribution of off-targets (read islands that are not assigned to RNA probes) and we selected $P < 0.1$ for expressed RNA probes (Supplemental Fig. S8A).

qRT-PCR

The tissue-specific expression profiles of 23 selected CRS regions (CaptureSeq $P < 0.1$) were determined by qRT-PCR using purified total RNA from 7 different tissues (brain, colon, heart, kidney, liver, small intestines and testis) in human and mouse. Human total RNA from these seven tissues (plus fetal brain) were ordered (Clontech). The same tissues were isolated from 30 day old male mice (Balbc/J), and total RNA was extracted using a modified miRNeasy protocol (Qiagen). See Supplemental Methods.

RNA structure probing

CRS sequences from human and mouse were selected for structure probing based on low FDRs, low SI between human and mouse, and expression in human and mouse brain as determined by qRT-PCR. Templates were made by PCR on gDNA templates and designed to include flanking sequences to the extent that it would facilitate the predicted structures upon folding (Hecker et al. 2015). CRS in vitro transcripts were screened by native gel electrophoresis and pairs that yielded single bands subjected to

structure probing using RNase V1 and S1 nuclease or Pb2+ for demonstration of double- and single-stranded regions, respectively. See Supplemental Methods.

Definition of gene regulatory regions

Initially, we defined enhancers through ENCODE chromatin segmentation states (classes E or WE in ≥2 of 6 human cell lines; length 100bp to 1kb) (Hoffman et al. 2013), loci upstream (-2kb to -100bp) of TSSs and loci downstream (+1bp to +2kb) of 3' ends of mRNAs and lncRNAs annotated in GENCODE. We considered only genes of >10kb away from their adjacent annotated genes (~13k genes). For more stringent definition of regulatory regions taking experimentally measured transcript boundaries into account we used the following data: DNase I Hypersensitivity Peak Clusters from ENCODE (95 cell types) ("The ENCODE Project Consortium" 2009), CAGE expression of robust (>10TPM) peaks (length ≤200bp) from FANTOM5 Phase 2.0 (Andersson et al. 2014a), ENCODE chromatin segmentation states, GENCODE gene/TSS annotation of mRNAs and lncRNAs, and polyadenylation signals (Gruber et al. 2016). TSS upstream transcription and enhancers were defined by CAGE peaks, DHSs and chromatin states. Alternative 3' ends were defined by poly(A) signals. See Supplemental Material and Supplemental Fig. S11.

Expression of structured and unstructured gene regulatory regions

Differences in expression level in tissues between structured (ones overlapping CRSs) and unstructured regulatory regions in different GC content or phastCons score bins were tested by one-sided Mann-Whitney *U* test, BH-corrected, and considered significant if $P < 0.05$ in all replicates. Uniquely mapped reads from ENCODE phase 3 and E-MTAB-513 (Supplemental Table S8) were counted for overlap with regulatory regions using featureCounts v1.5.0-p1 (parameters -s 0 -T 8 -Q 10 -p -P -d 50 -D 2000 -C -B --read2pos 5). Read counts were normalized to CPM/RLE before hypothesis testing. GC content and phastCons were measured from regulatory regions.

Transcript stability

To test for stability of transcripts we analyzed triplicate published CAGE libraries from hRRP40 (exosome) depleted and EGFP depleted (control) HeLa cells (Andersson et al. 2014b). We got premapped 5' ends of sequenced capped RNAs for all 6 libraries from the authors (.bed). Exosome sensitivity was calculated as described in (Andersson et al. 2014b) for both strands by $\max((E_{exo} - E_{ctr})/E_{exo}, 0)$, with E_{exo} and E_{ctr} denoting the expression level after exosome depletion and in control HeLa cells, respectively. We used thresholds of ≤0.25 and ≥0.75 to identify highly stable and highly unstable RNAs emanating from transcribed DHSs. See Supplemental Methods.

Statistics and visualization

Statistical analyses and visualization were performed with R (R Core Team 2016), feature distances were calculated using BEDTools (Quinlan and Hall 2010), genomic views were from UCSC Genome Browser (Rosenbloom et al. 2015), and structures/alignments were drawn with Vienna RNA Package tools (Lorenz et al. 2011).

Data access

The CaptureSeq data has been submitted to the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE87214. The list of CRSs, CRS alignments, CRS annotation, and on-target expressed (P<0.1) regions found by CaptureSeq are provided as Supplemental Data. Our catalog of predicted CRSs, Supplemental Data and a UCSC track hub of CRSs are available at http://rth.dk/resources/rnannotator/crs/vert.

Acknowledgements

We would like to thank Christian Anthon for computational and infrastructural support, Sachin Pundhir for discussion on the RNA-seq data, Robin Andersson and Albin Sandelin for sharing raw RNA-seq data of hRRP40 depleted and EGFP depleted HeLa cells. We thank Rasmus Nielsen, Torben Heick Jensen, Lars Juhl Jensen and Peter F Stadler for fruitful discussions and suggestions, and the anonymous reviewers for their helpful and constructive comments. This work was supported by Innovation Fund Denmark (0603-00320B), Danish Center for Scientific Computing (DCSC, DeIC), and The Lundbeck Foundation (R19-A2306).

Figure 1: Performance assessment, genomic distribution and conservation of CRS predictions. (*A*) Mean FDR of CRSs for different CMfinder score (pscore) cutoffs and GC content intervals. FDR calculation is based on SISSIz (Gesell and Washietl 2008) simulated alignments. The large decrease in FDR observed between pscore cutoff 40 and 50 motivated us to base all further analyses on pscore≥50. The mean FDR covering all ranges of GC content is 15.8. (*B*) GC content of CRS region alignments. (*C*) Fold enrichment of CRSs regions for biotypes and previous computational RNA structure screens in vertebrates (blue). (*D*) Absolute CRS region coverage of biotypes. (*E*) Relative position of CRS regions over noncoding biotypes presented as fold enrichment of CRS regions in bins, each 5% (considering only exons) of the feature's (UTR or gene) length. The trend of decreasing number of structures from 5' to 3' is common to 5' UTRs and lncRNAs. *(F)* Number of CRSs conserved in the 100-species tree. *(G)* Average pairwise sequence identity (SI) of CRS region alignments over the 17 representative genomes in the phylogenetic tree. *(H)* Re-alignment (calculated as in (Torarinsson et al. 2008)) compares the 17-species MULTIZ alignment blocks (hg18) to corresponding structure-based alignments of CRS regions (17-way subalignments extracted from our 100-species/hg38 results, as described in Methods) . (*I*) Species number of CRS region alignments. In (*B*), (*G*) and (*I*) the CRSs of highest GC content, SI and species number, respectively, are used as representatives of a CRS region, and in (*H*) the CRSs of lowest re-alignment are used as representatives. The biotypes in (*G*), (*H*) and (*I*) are sorted by their median SI.

Figure 2: Human and mouse conservation of CRS regions is reflected by binding sites of RBPs and expression. (*A*) 7 of 10 RBPs display enrichment of CRSs in conserved binding sites ($P < 10^{-7}$, FET). Significant enrichments are colored dark blue; light blue were not significant. (*B*) A relatively large number of CRSs (146,670) are expressed in both human and mouse (red bars) over four tissues (heart, liver, diencephalon/forebrain and cerebellum/hindbrain) with comparable total RNA-seq data (Methods). In total 157,136 CRSs are expressed in both human and mouse in total RNA-seq or poly(A) RNA-seq (Supplemental Fig. S7). CRSs with an empirical *P*-value < 0.01 were assigned an "expressed" state. We considered only 433,327 of 543,390 human-mouse conserved CRSs which have the same biotype in both species. Note that "5' extension" and "3' extension" refer to 2kb regions up- and down-stream of UTRs and lncRNAs; UTRs themselves are included in "mRNA". (*C*) Expression correlation between human and mouse for different biotypes was measured by Pearson's correlation coefficient *r* of expression levels in poly(A) RNA-seq (six tissues: testis, liver, kidney, heart, cerebellum and brain). "Background" is sampled over the input MA blocks with human-mouse conservation not overlapping the other biotypes. The number on the left of violin plots is total number of measured CRSs with expression in at least two tissues, and the number on the right side is number of CRSs with $r > 0.8$.

Figure 3: CaptureSeq and qRT-PCR show conserved expression of CRSs. (*A*) ROC curve of CRS region detection in brain based on public poly(A) RNA-seq defined by different CPM/RLE cutoffs (numbers on the curves) using the CRS region detection through CaptureSeq in fetal brain as gold standard. (*B*) Expression profiles of 23 CRS regions were measured with qRT-PCR (normalized by CRS regions) in 7 tissues in both human and mouse. The CRS regions have weakly conserved primary sequences and were expressed in the CaptureSeq ($P < 0.1$). The CRS regions are sorted by decreasing Pearson's correlation coefficients of expression profiles between human and mouse. (*C*) The CRS region C3381920 is located in the 3' end of the lncRNA *AC07304.25*. Despite of no expression in brain in publicly available total and poly(A) RNA-seq data, it showed up in human brain in both CaptureSeq and qRT-PCR. Common expression in human and mouse was observed in the gastrointestinal tract (small intestine and colon; see (*B*)). Region C3381920 contains the CRS M0653745 whose structure is highly conserved between human and mouse. Color code in human and mouse structures is base pair probabilities calculated by the Vienna RNA package (Lorenz et al. 2011).

Figure 4: *In vitro* RNA structure probing in human and mouse shows conserved structure of CRS M1695693. FDR is 11.0% and SI of the 9-species (filtered from the 17-species tree) structural alignment is 48% (45% between human and mouse). The CRS is located between the 3' UTRs of *HOMER2* (minus strand) and *WHAMM* (plus strand) (chr15:8284671-82846804). It overlaps a DHS (ENCODE) and has the typical chromatin signatures of enhancers, namely enrichment of H3K4me1 and reduced enrichment of H3K4me3, all indicators for a transcribed regulatory region. However, CAGE data from FANTOM5 did not support this hypothesis, instead, poly(A) site clusters (Gruber et al. 2016) suggest an extended 3' UTR of *HOMER2*. (*A*) Genomic tracks. (*B*) Structure probing results in human and mouse where red marks base-paired nucleotides (ds), and green and blue mark single-stranded nucleotides (ss). (*C*) CMfinder's structural alignment, predicted consensus RNA secondary structure, and predicted individual structures in human and mouse as dot-bracket notation. The probing results are overlapped with the *in-silico* predictions by their color code.

Figure 5: Coverage and expression of CRS regions in gene regulatory regions. The figure's three rows describe regions surrounding (i) enhancers (panels *A-D*); (ii) most distal TSS of mRNAs/lncRNAs (panels *E-H*); and (iii) most distal 3' end of mRNAs/lncRNAs (panels *I-K*), respectively. Panels (*A,E,I*) plot density of CRS regions near those features: counts in 50bp windows normalized by the number of features. "Predicted" curves (orange) reflect all CRS regions; "Transcribed" curves (blue) reflect the subset supported by unannotated transcription boundaries. Lower subpanels show estimated FDRs (mean, standard deviation) of those predictions. All other panels are based on the "transcribed" subset; for details see Methods "Definition of gene regulatory regions" and Supplemental Fig. S11. In summary, expression is based on: (*B,C*) CAGE TSS near enhancers, *(F,G)* CAGE TSS upstream antisense w.r.t. mRNA/lncRNA, (*J,K*) active poly(A) sites downstream sense w.r.t. mRNA/lncRNA. "Structured"/"CRS" denote regions that overlap CRSs; "unstructured"/"no CRS" do not. *(B,C,F,G)* Total RNA-seq in fetal human cerebellum (technical replicate two of experiment ENCSR000AEW; ENCODE Phase 3). *(J,K)* Poly(A) RNA-seq of human brain (HBM). *(B,F,J)* Expression levels are in counts per million after cross-experiment relative log expression normalization (CPM/RLE). *(C,G,K)* GC content and phastCons (from 100-species MULTIZ alignments) of expressed structured (CRS) versus unstructured regions (no CRS). Expressed regions were defined by empirical *P*-value < 0.01 and CPM/RLE$\geq$1. (*D,H*) Transcript stability at ENCODE HeLa DHSs, as described in (Andersson et al. 2014b), and GC content of structured (CRS) and unstructured regions (no CRS). Odds ratios quantify how strongly stability is associated with CRS overlap.

Figure 6: Example CRSs in gene regulatory regions are supported by unannotated transcript boundaries. (*A*) Two intergenic enhancers in a highly structured region of low SI (CRS M1293227 is only conserved in primates) between two gene 3' ends. (*B*) *MIR320A* is upstream of *POLR3D* TSS. (*C*) Antisense transcription at the promoter of the lncRNA *LINC01132* has enhancer-like chromatin signatures. (*D*) Intergenic enhancer with unidirectional stable transcription from the minus strand as measured by control and exosome-depleted HeLa cells (Andersson et al. 2014b). Color code in consensus structures is the level of base pair conservation in the structure-based alignments.

Reference List

"The ENCODE Project Consortium". 2009. Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. *Nature* **457**:1028-1032

"The ENCODE Project Consortium". 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**:57-74

Almada, A.E., Wu, X., Kriz, A.J., Burge, C.B., and Sharp, P.A. 2013. Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature* **499**:360-363

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al. 2014a. An atlas of active enhancers across human cell types and tissues. *Nature* **507**:455-461

Andersson, R., Refsing, A.P., Valen, E., Core, L.J., Bornholdt, J., Boyd, M., Heick, J.T., and Sandelin, A. 2014b. Nuclear stability and transcriptional directionality separate functionally distinct RNA species. *Nat.Commun.* **5**:5336

Arner, E., Daub, C.O., Vitting-Seerup, K., Andersson, R., Lilje, B., Drablos, F., Lennartsson, A., Ronnerblad, M., Hrydziuszko, O., Vitezic, M., et al. 2015. Gene regulation. Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science* **347**:1010-1014

Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. 2015. A global reference for human genetic variation. *Nature* **526**:68-74

Aw, J.G., Shen, Y., Wilm, A., Sun, M., Lim, X.N., Boon, K.L., Tapsin, S., Chan, Y.S., Tan, C.P., Sim, A.Y., et al. 2016. In Vivo Mapping of Eukaryotic RNA Interactomes Reveals Principles of Higher-Order Organization and Regulation. *Mol.Cell* **62**:603-617

Berkovits, B.D. and Mayr, C. 2015. Alternative 3' UTRs act as scaffolds to regulate membrane protein localization. *Nature* **522**:363-367

Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F.A., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D., et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Research* **14**:708-715

Burge, S.W., Daub, J., Eberhardt, R., Tate, J., Barquist, L., Nawrocki, E.P., Eddy, S.R., Gardner, P.P., and Bateman, A. 2013. Rfam 11.0: 10 years of RNA families. *Nucleic Acids Research* **41**:D226-D232

Di Giammartino, D.C., Nishida, K., and Manley, J.L. 2011. Mechanisms and consequences of alternative polyadenylation. *Mol.Cell* **43**:853-866

Ding, Y., Tang, Y., Kwok, C.K., Zhang, Y., Bevilacqua, P.C., and Assmann, S.M. 2014. In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature* **505**:696-700

Elkon, R., Ugalde, A.P., and Agami, R. 2013. Alternative cleavage and polyadenylation: extent, regulation and function. *Nat.Rev.Genet.* **14**:496-506

Fang, R., Moss, W.N., Rutenberg-Schoenberg, M., and Simon, M.D. 2015. Probing Xist RNA Structure in Cells Using Targeted Structure-Seq. *PLoS.Genet.* **11**:e1005668

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J.Mol.Evol.* **17**:368-376

Gesell, T. and Washietl, S. 2008. Dinucleotide controlled null models for comparative RNA gene prediction. *BMC Bioinformatics* **9**:248

Glover-Cutter, K., Kim, S., Espinosa, J., and Bentley, D.L. 2008. RNA polymerase II pauses and associates with pre-mRNA processing factors at both ends of genes. *Nat.Struct.Mol.Biol.* **15**:71-78

Goodarzi, H., Najafabadi, H.S., Oikonomou, P., Greco, T.M., Fish, L., Salavati, R., Cristea, I.M., and Tavazoie, S. 2012. Systematic discovery of structural elements governing stability of mammalian messenger RNAs. *Nature* **485**:264-268

Gorodkin, J., Hofacker, I.L., Torarinsson, E., Yao, Z., Havgaard, J.H., and Ruzzo, W.L. 2010. De novo prediction of structured RNAs from genomic sequences. *Trends Biotechnol.* **28**:9-19

Gruber, A.J., Schmidt, R., Gruber, A.R., Martin, G., Ghosh, S., Belmadani, M., Keller, W., and Zavolan, M. 2016. A comprehensive analysis of 3' end sequencing data sets reveals novel polyadenylation signals and the repressive role of heterogeneous ribonucleoprotein C on cleavage and polyadenylation. *Genome Res.* **26**:1145-1159

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., et al. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**:1760-1774

Hecker, N., Christensen-Dalsgaard, M., Seemann, S.E., Havgaard, J.H., Stadler, P.F., Hofacker, I.L., Nielsen, H., and Gorodkin, J. 2015. Optimizing RNA structures by sequence extensions using RNAcop. *Nucleic Acids Res.* **43**:8135-8145

Hezroni, H., Koppstein, D., Schwartz, M.G., Avrutin, A., Bartel, D.P., and Ulitsky, I. 2015. Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Rep.* **11**:1110-1122

Hoffman, M.M., Ernst, J., Wilder, S.P., Kundaje, A., Harris, R.S., Libbrecht, M., Giardine, B., Ellenbogen, P.M., Bilmes, J.A., Birney, E., et al. 2013. Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Research* **41**:827-841

Jensen, T.H., Jacquier, A., and Libri, D. 2013. Dealing with pervasive transcription. *Mol.Cell* **52**:473-484

Kim, D.H., Saetrom, P., Snove, O., Jr., and Rossi, J.J. 2008. MicroRNA-directed transcriptional gene silencing in mammalian cells. *Proc.Natl.Acad.Sci.U.S.A* **105**:16230-16235

Kim, Y.K., Furic, L., Desgroseillers, L., and Maquat, L.E. 2005. Mammalian Staufen1 recruits Upf1 to specific mRNA 3'UTRs so as to elicit mRNA decay. *Cell* **120**:195-208

Li, W., Notani, D., Ma, Q., Tanasa, B., Nunez, E., Chen, A.Y., Merkurjev, D., Zhang, J., Ohgi, K., Song, X., et al. 2013. Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature* **498**:516-520

Liao, Y., Smyth, G.K., and Shi, W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics.* **30**:923-930

Lin, M.F., Jungreis, I., and Kellis, M. 2011. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics.* **27**:i275-i282

Lin, S., Lin, Y., Nery, J.R., Urich, M.A., Breschi, A., Davis, C.A., Dobin, A., Zaleski, C., Beer, M.A., Chapman, W.C., et al. 2014. Comparison of the transcriptional landscapes between human and mouse tissues. *Proc.Natl.Acad.Sci.U.S.A* **111**:17224-17229

Lorenz, R., Bernhart, S.H., Honer Zu, S.C., Tafer, H., Flamm, C., Stadler, P.F., and Hofacker, I.L. 2011. ViennaRNA Package 2.0. *Algorithms.Mol.Biol.* **6**:26

Lu, Z., Zhang, Q.C., Lee, B., Flynn, R.A., Smith, M.A., Robinson, J.T., Davidovich, C., Gooding, A.R., Goodrich, K.J., Mattick, J.S., et al. 2016. RNA Duplex Map in Living Cells Reveals Higher-Order Transcriptome Structure. *Cell* **165**:1267-1279

Lunde, B.M., Moore, C., and Varani, G. 2007. RNA-binding proteins: modular design for efficient function. *Nat.Rev.Mol.Cell Biol.* **8**:479-490

Lunter, G., Ponting, C.P., and Hein, J. 2006. Genome-wide identification of human functional DNA using a neutral indel model. *Plos Computational Biology* **2**:2-12

Macias, S., Plass, M., Stajuda, A., Michlewski, G., Eyras, E., and Caceres, J.F. 2012. DGCR8 HITS-CLIP reveals novel functions for the Microprocessor. *Nat.Struct.Mol.Biol.* **19**:760-766

Managadze, D., Rogozin, I.B., Chernikova, D., Shabalina, S.A., and Koonin, E.V. 2011. Negative correlation between expression level and evolutionary rate of long intergenic noncoding RNAs. *Genome Biol.Evol.* **3**:1390-1404

Mercer, T.R., Clark, M.B., Crawford, J., Brunck, M.E., Gerhardt, D.J., Taft, R.J., Nielsen, L.K., Dinger, M.E., and Mattick, J.S. 2014. Targeted sequencing for gene discovery and quantification using RNA CaptureSeq. *Nat.Protoc.* **9**:989-1009

Miladi, M., Junge, A., Costa, F., Seemann, S.E., Hull, H.J., Gorodkin, J., and Backofen, R. 2017. RNAscClust: clustering RNA sequences using structure conservation and graph based motifs. *Bioinformatics.*

Nawrocki, E.P., Burge, S.W., Bateman, A., Daub, J., Eberhardt, R.Y., Eddy, S.R., Floden, E.W., Gardner, P.P., Jones, T.A., Tate, J., et al. 2015. Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.* **43**:D130-D137

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

Ntini, E., Jarvelin, A.I., Bornholdt, J., Chen, Y., Boyd, M., Jorgensen, M., Andersson, R., Hoof, I., Schein, A., Andersen, P.R., et al. 2013. Polyadenylation site-induced decay of upstream transcripts enforces promoter directionality. *Nat.Struct.Mol.Biol.* **20**:923-928

Parker, B.J., Moltke, I., Roth, A., Washietl, S., Wen, J., Kellis, M., Breaker, R., and Pedersen, J.S. 2011. New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. *Genome Res.* **21**:1929-1943

Pedersen, J.S., Bejerano, G., Siepel, A., Rosenbloom, K., Lindblad-Toh, K., Lander, E.S., Kent, J., Miller, W., and Haussler, D. 2006. Identification and classification of conserved RNA secondary structures in the human genome. *Plos Computational Biology* **2**:251-262

Ponjavic, J., Ponting, C.P., and Lunter, G. 2007. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res.* **17**:556-565

Ponting, C.P. 2008. The functional repertoires of metazoan genomes. *Nat.Rev.Genet.* **9**:689-698

Preker, P., Almvig, K., Christensen, M.S., Valen, E., Mapendano, C.K., Sandelin, A., and Jensen, T.H. 2011. PROMoter uPstream Transcripts share characteristics with mRNAs and are produced upstream of all three major types of mammalian promoters. *Nucleic Acids Res.* **39**:7179-7193

Quinlan, A.R. and Hall, I.M. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**:841-842

R Core Team. R: A Language and Environment for Statistical Computing. 2016. Vienna, Austria, R Foundation for Statistical Computing.

Ref Type: Computer Program

Rands, C.M., Meader, S., Ponting, C.P., and Lunter, G. 2014. 8.2% of the Human Genome Is Constrained: Variation in Rates of Turnover across Functional Element Classes in the Human Lineage. *Plos Genetics* **10**

Ray, D., Kazan, H., Cook, K.B., Weirauch, M.T., Najafabadi, H.S., Li, X., Gueroussov, S., Albu, M., Zheng, H., Yang, A., et al. 2013. A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**:172-177

Rinn, J.L. and Chang, H.Y. 2012. Genome regulation by long noncoding RNAs. *Annu.Rev.Biochem.* **81**:145-166

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.* **26**:139-140

Rosenbloom, K.R., Armstrong, J., Barber, G.P., Casper, J., Clawson, H., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haeussler, M., et al. 2015. The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res.* **43**:D670-D681

Rouskin, S., Zubradt, M., Washietl, S., Kellis, M., and Weissman, J.S. 2014. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature* **505**:701-705

Rybak-Wolf, A., Jens, M., Murakawa, Y., Herzog, M., Landthaler, M., and Rajewsky, N. 2014. A variety of dicer substrates in human and C. elegans. *Cell* **159**:1153-1167

Seila, A.C., Calabrese, J.M., Levine, S.S., Yeo, G.W., Rahl, P.B., Flynn, R.A., Young, R.A., and Sharp, P.A. 2008. Divergent Transcription from Active Promoters. *Science* **322**:1849-1851

Sharma, E., Sterne-Weiler, T., O'Hanlon, D., and Blencowe, B.J. 2016. Global Mapping of Human RNA-RNA Interactions. *Mol.Cell* **62**:618-626

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M.M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research* **15**:1034-1050

Smit, AFA, Hubley, R, and Green, P. http://www.repeatmasker.org . 2013.

Ref Type: Online Source

Smith, M.A., Gesell, T., Stadler, P.F., and Mattick, J.S. 2013. Widespread purifying selection on RNA structure in mammals. *Nucleic Acids Res.* **41**:8220-8236

Sundfeld, D., Havgaard, J.H., de Melo, A.C., and Gorodkin, J. 2016. Foldalign 2.5: multithreaded implementation for pairwise structural RNA alignment. *Bioinformatics.* **32**:1238-1240

Torarinsson, E., Yao, Z., Wiklund, E.D., Bramsen, J.B., Hansen, C., Kjems, J., Tommerup, N., Ruzzo, W.L., and Gorodkin, J. 2008. Comparative genomics beyond sequence-based alignments: RNA structures in the ENCODE regions. *Genome Res.* **18**:242-251

Ulitsky, I. and Bartel, D.P. 2013. lincRNAs: Genomics, Evolution, and Mechanisms. *Cell* **154**:26-46

Wan, Y., Qu, K., Zhang, Q.C., Flynn, R.A., Manor, O., Ouyang, Z., Zhang, J., Spitale, R.C., Snyder, M.P., Segal, E., et al. 2014. Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* **505**:706-709

Wang, A.X., Ruzzo, W.L., and Tompa, M. 2007. How accurately is ncRNA aligned within whole-genome multiple alignments? *BMC.Bioinformatics.* **8**:417

Washietl, S., Hofacker, I.L., Lukasser, M., Huttenhofer, A., and Stadler, P.F. 2005. Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat.Biotechnol.* **23**:1383-1390

Washietl, S., Kellis, M., and Garber, M. 2014. Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. *Genome Research* **24**:616-628

Waters, L.S. and Storz, G. 2009. Regulatory RNAs in bacteria. *Cell* **136**:615-628

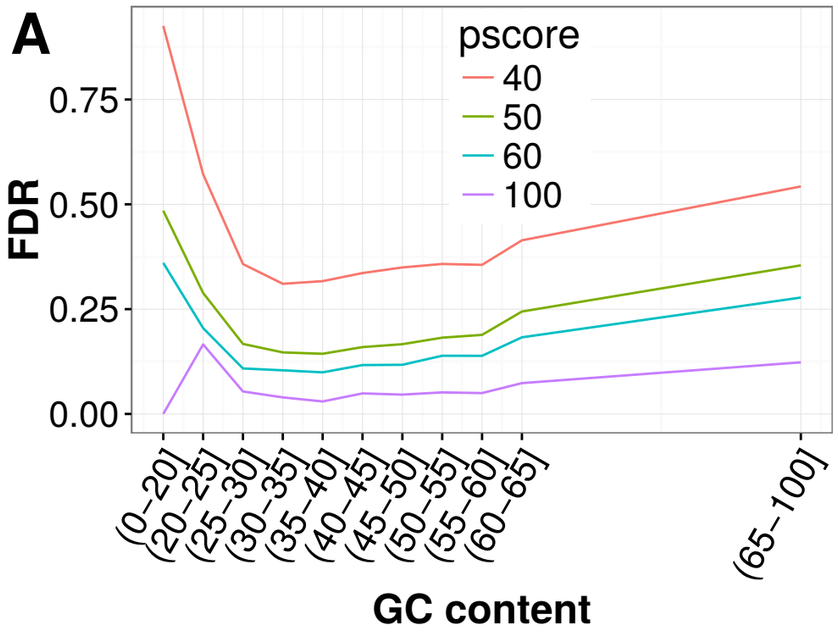
Weinberg, Z., Barrick, J.E., Yao, Z., Roth, A., Kim, J.N., Gore, J., Wang, J.X., Lee, E.R., Block, K.F., Sudarsan, N., et al. 2007. Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline. *Nucleic Acids Res.* **35**:4809-4819

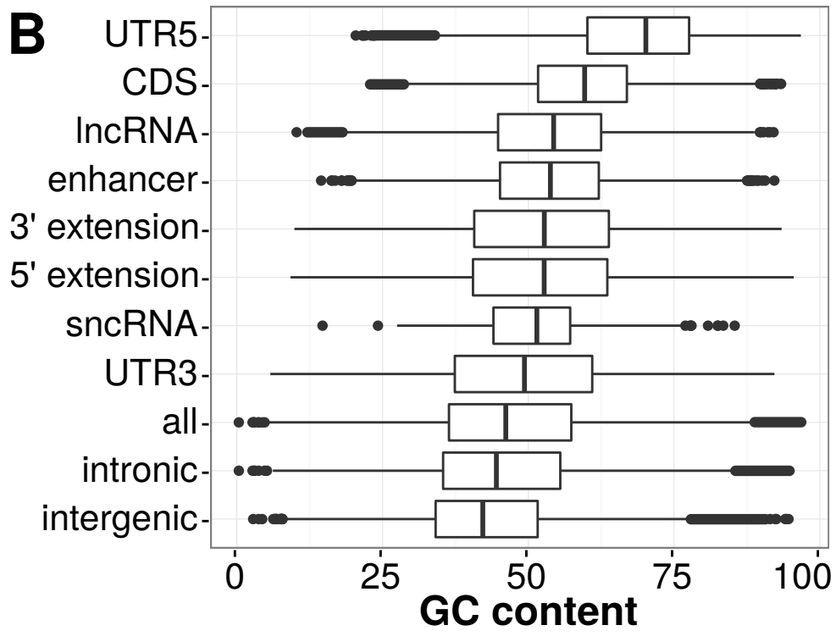
Weinberg, Z., Perreault, J., Meyer, M.M., and Breaker, R.R. 2009. Exceptional structured noncoding RNAs revealed by bacterial metagenome analysis. *Nature* **462**:656-659

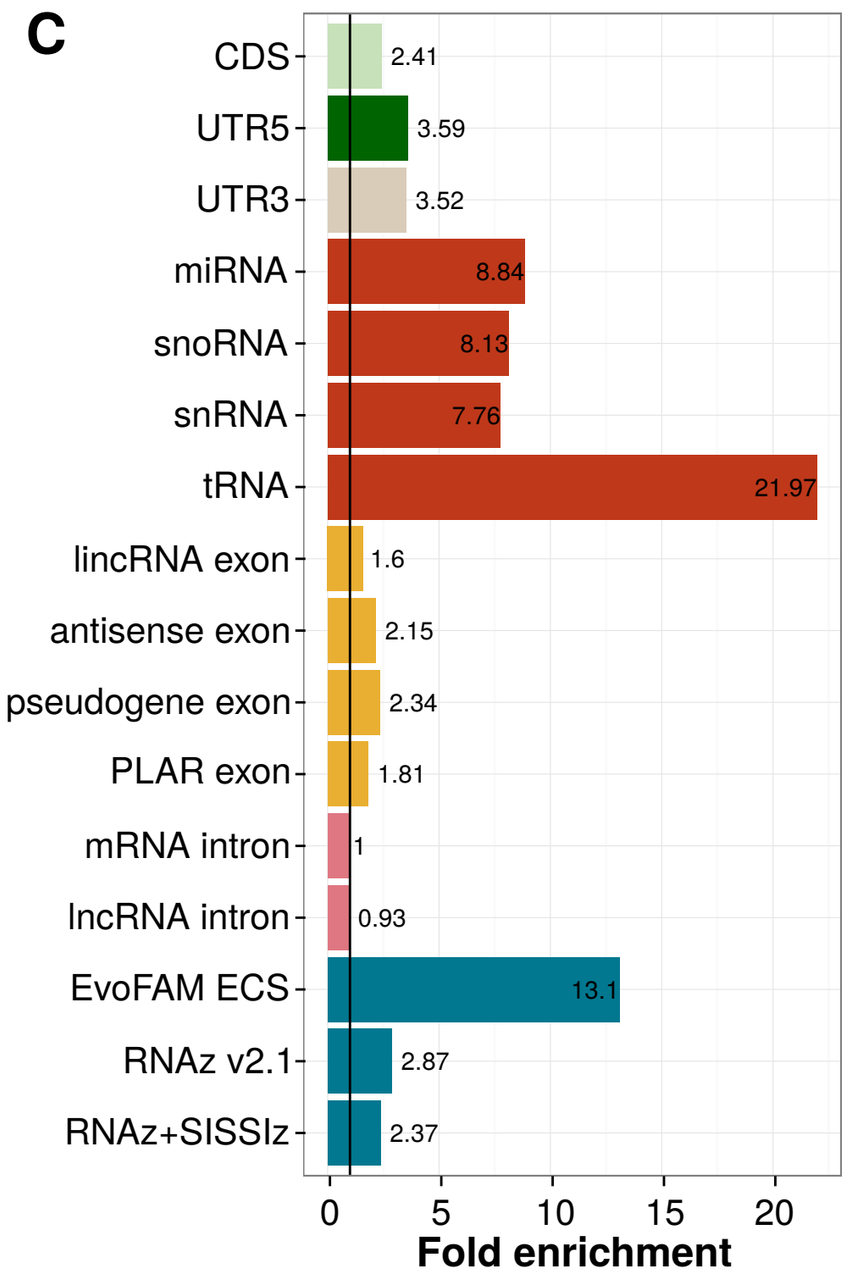
Yao, Z. 2008. Genome Scale Search of Noncoding RNAs: Bacteria to Vertebrates. *PhD Thesis*:(http://homes.cs.washington.edu/~ruzzo/theses/zizhen-yao/zizhen_uwthesis.pdf)

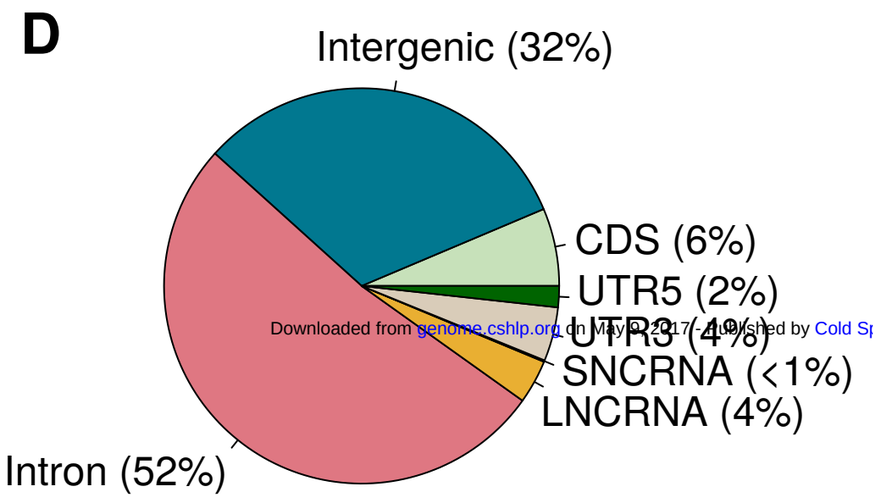
Yao, Z., Weinberg, Z., and Ruzzo, W.L. 2006. CMfinder-a covariance model based RNA motif finding algorithm. *Bioinformatics.* **22**:445-452

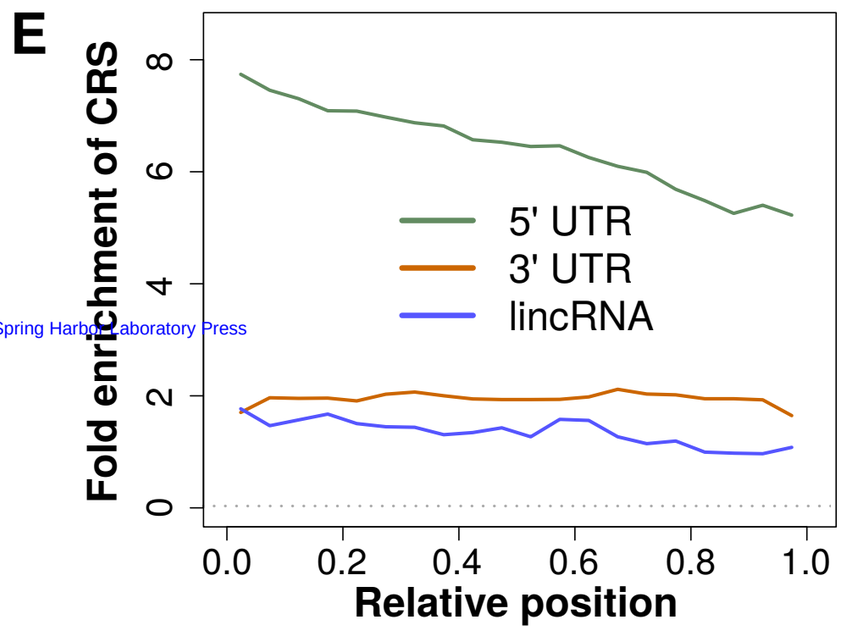
Zhang, Y., Yang, L., and Chen, L.L. 2014. Life without A tail: New formats of long noncoding RNAs. *Int.J.Biochem.Cell Biol.* **54C**:338-349

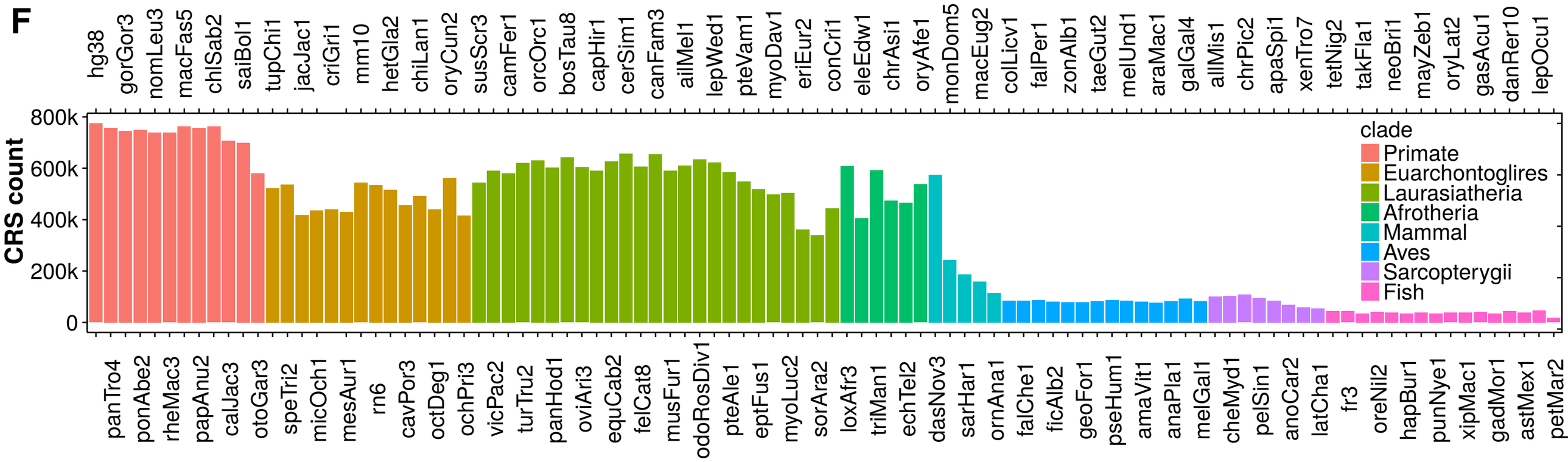


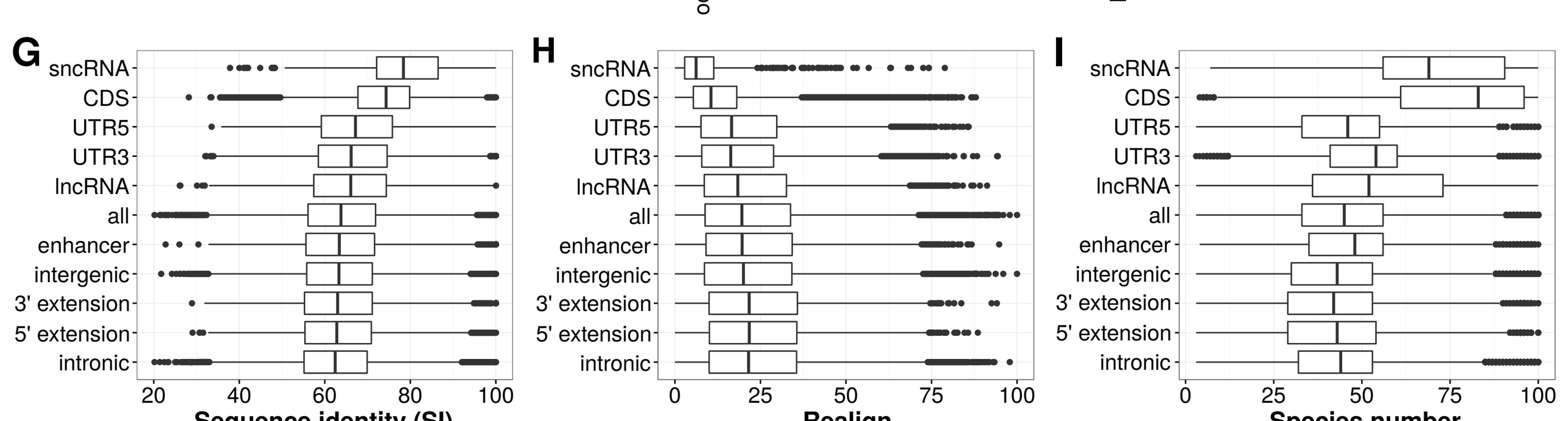


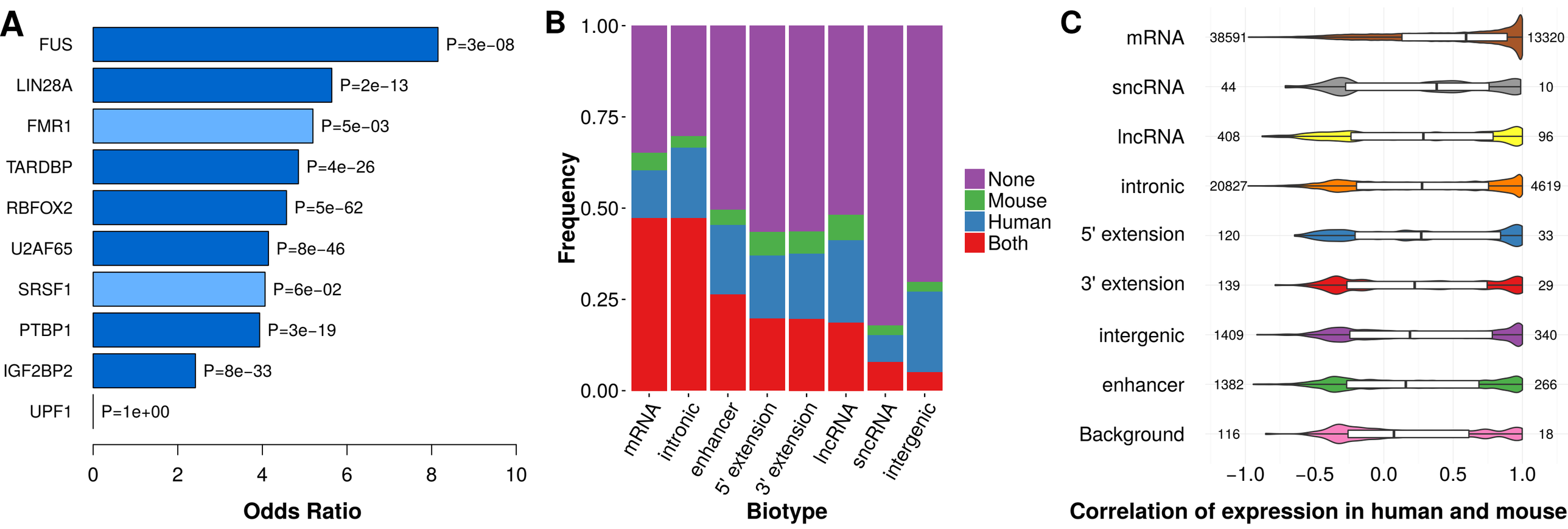


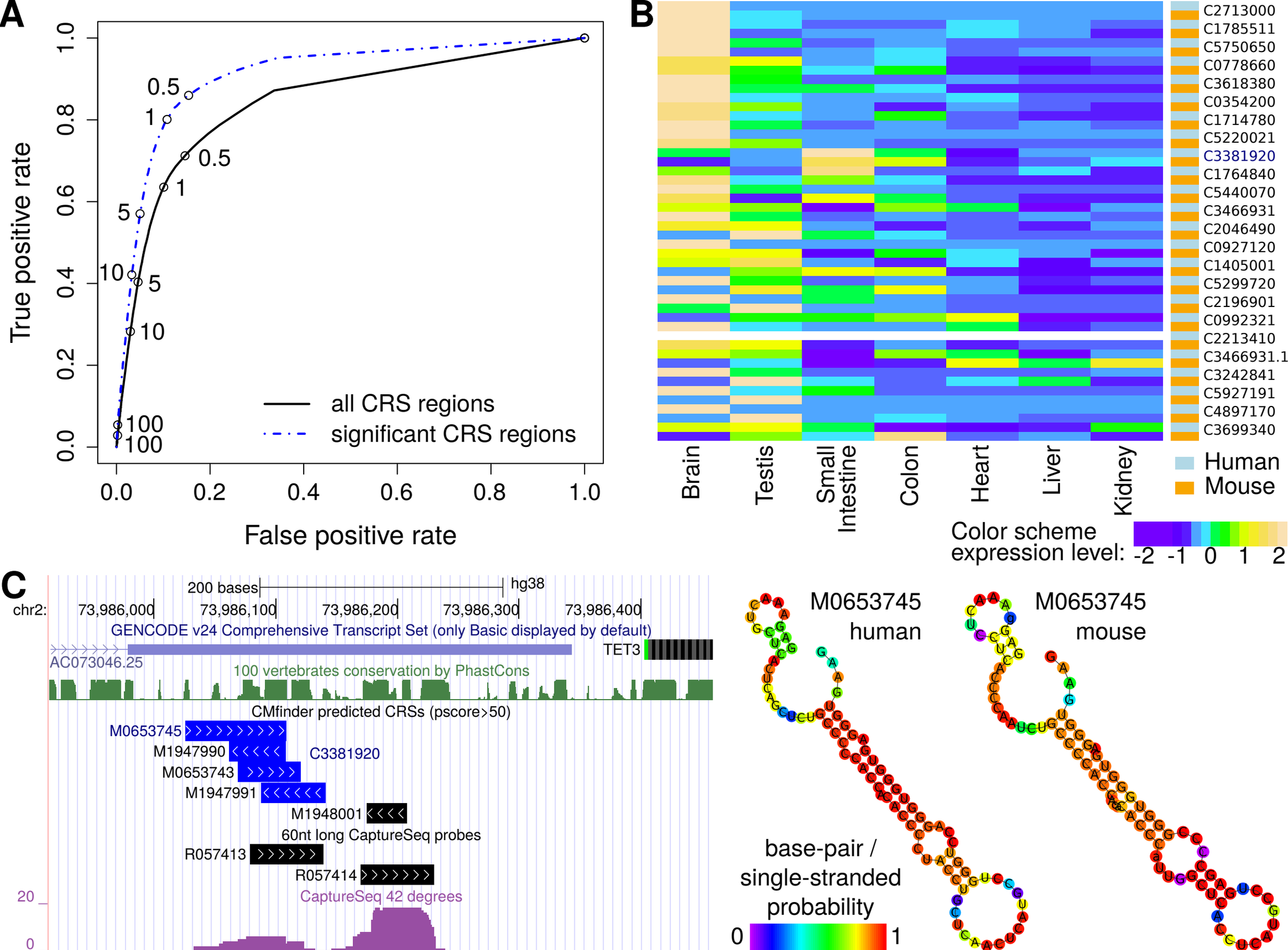




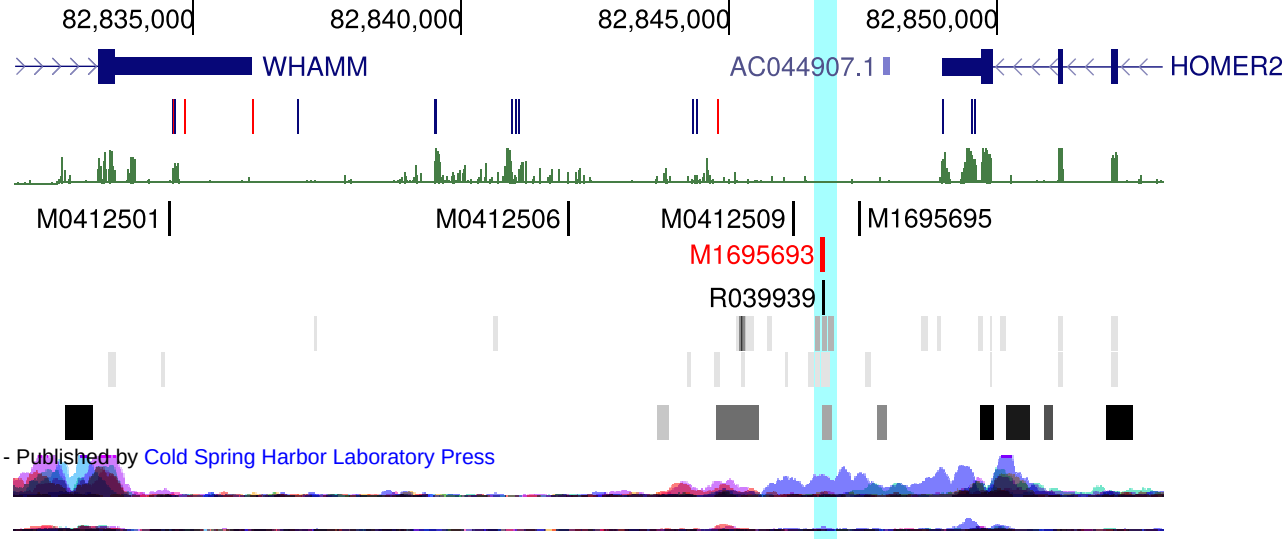




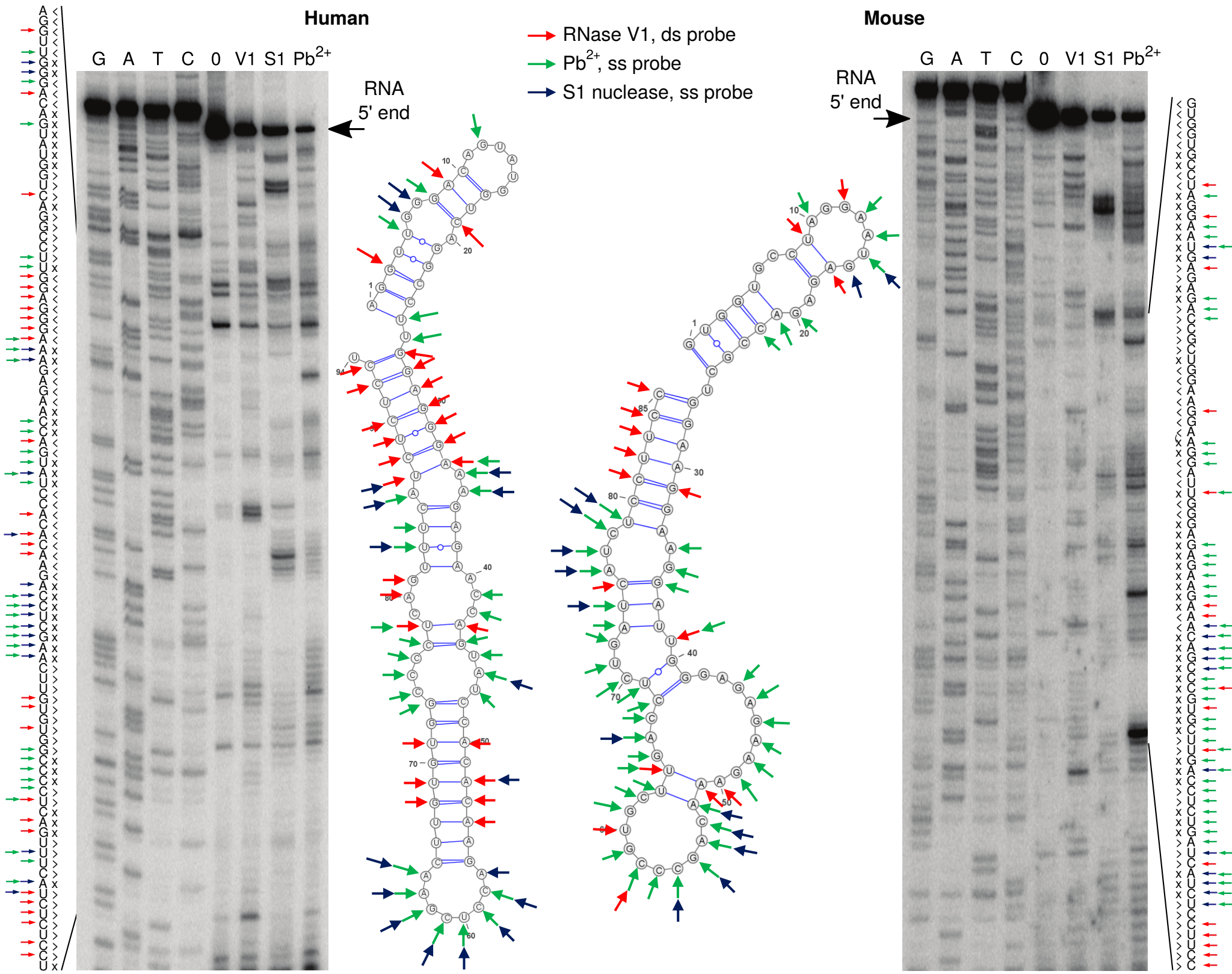

A
FUS
LIN28A
FMR1
TARDBP
RBFOX2
U2AF65
SRSF1
PTBP1
IGF2BP2
UPF1
P=3e−08
P=2e−13
P=5e−03
P=4e−26
P=5e−62
P=8e−46
P=6e−02
P=3e−19
P=8e−33
P=1e+00
0
2
4
6
8
10
Odds Ratio
B
Frequency
1.00
0.75
0.50
0.25
0.00
mRNA
intronic
enhancer
5' extension
3' extension
lncRNA
sncRNA
intergenic
Biotype
None
Mouse
Human
Both
C
mRNA
sncRNA
lncRNA
intronic
5' extension
3' extension
intergenic
enhancer
Background
38591
13320
44
10
408
96
20827
4619
120
33
139
29
1409
340
1382
266
116
18
−1.0
−0.5
0.0
0.5
1.0
Correlation of expression in human and mouse


A
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
0.5
1
0.5
1
5
10
5
10
100
100
all CRS regions
significant CRS regions
B
C2713000
C1785511
C5750650
C0778660
C3618380
C0354200
C1714780
C5220021
C3381920
C1764840
C5440070
C3466931
C2046490
C0927120
C1405001
C5299720
C2196901
C0992321
C2213410
C3466931.1
C3242841
C5927191
C4897170
C3699340
Brain
Testis
Small Intestine
Colon
Heart
Liver
Kidney
Human
Mouse
Color scheme expression level: -2 -1 0 1 2
C
200 bases
hg38
chr2:
73,986,000
73,986,100
73,986,200
73,986,300
73,986,400
GENCODE v24 Comprehensive Transcript Set (only Basic displayed by default)
AC073046.25
TET3
100 vertebrates conservation by PhastCons
CMfinder predicted CRSs (pscore>50)
M0653745
M1947990
C3381920
M0653743
M1947991
M1948001
60nt long CaptureSeq probes
R057413
R057414
CaptureSeq 42 degrees
20
0
M0653745 human
M0653745 mouse
base-pair / single-stranded probability
0
1

A



B



C

CRS region distribution

Expression

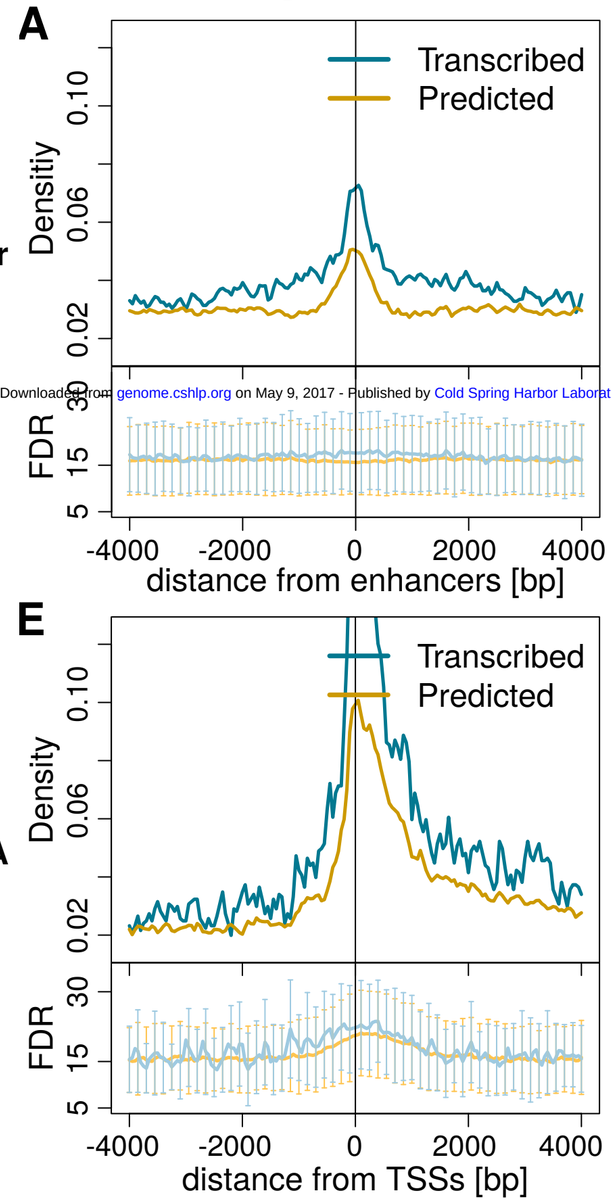
GC content/Conservation

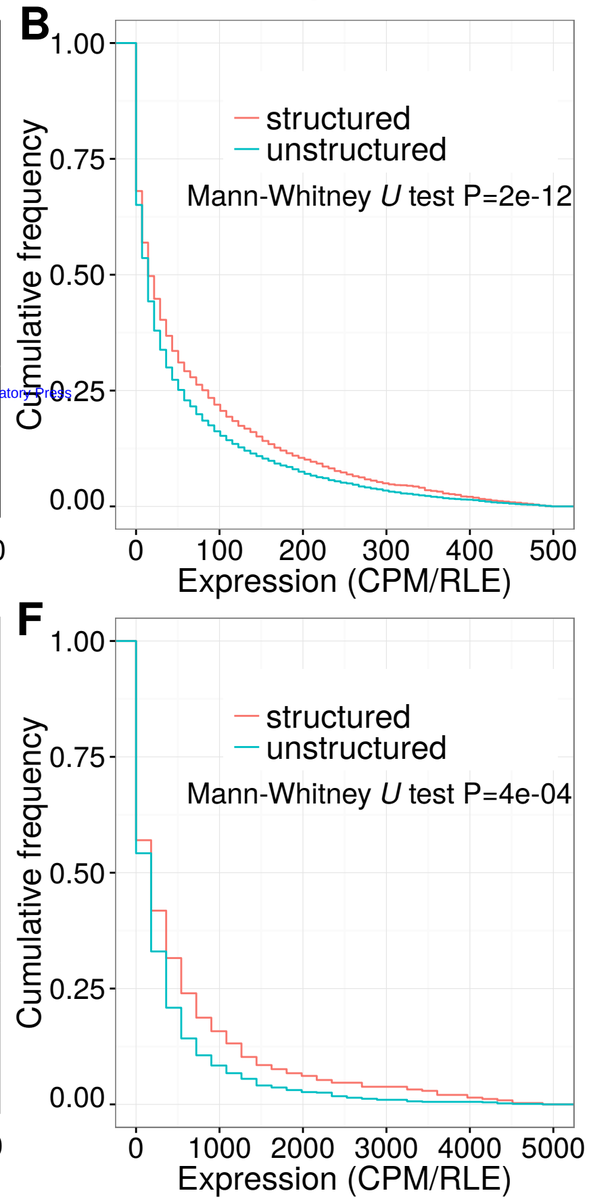
Transcript stability

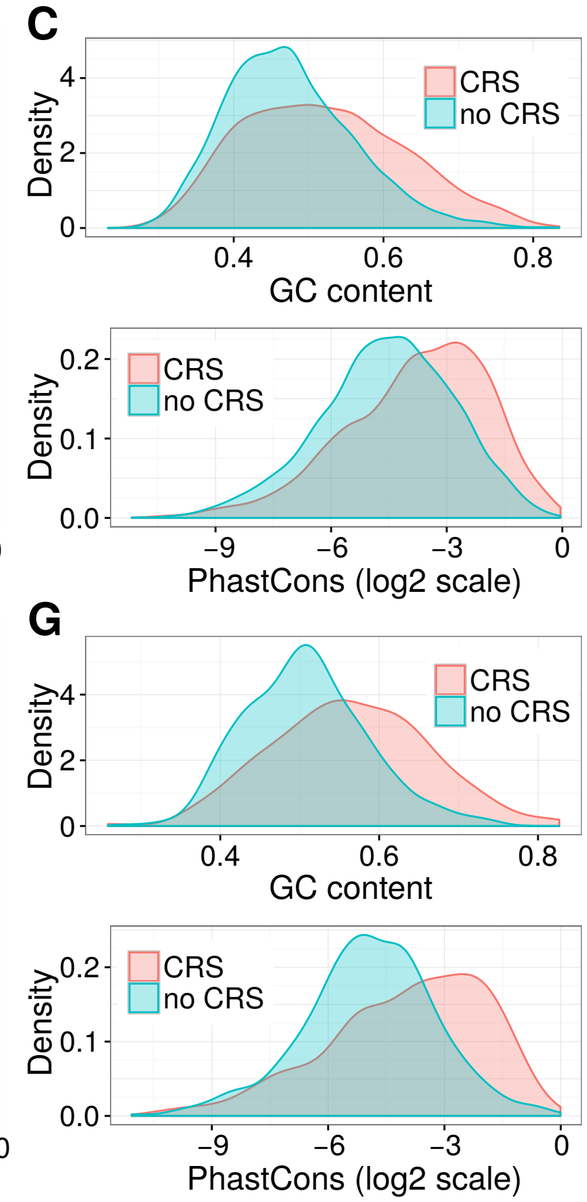
Enhancer

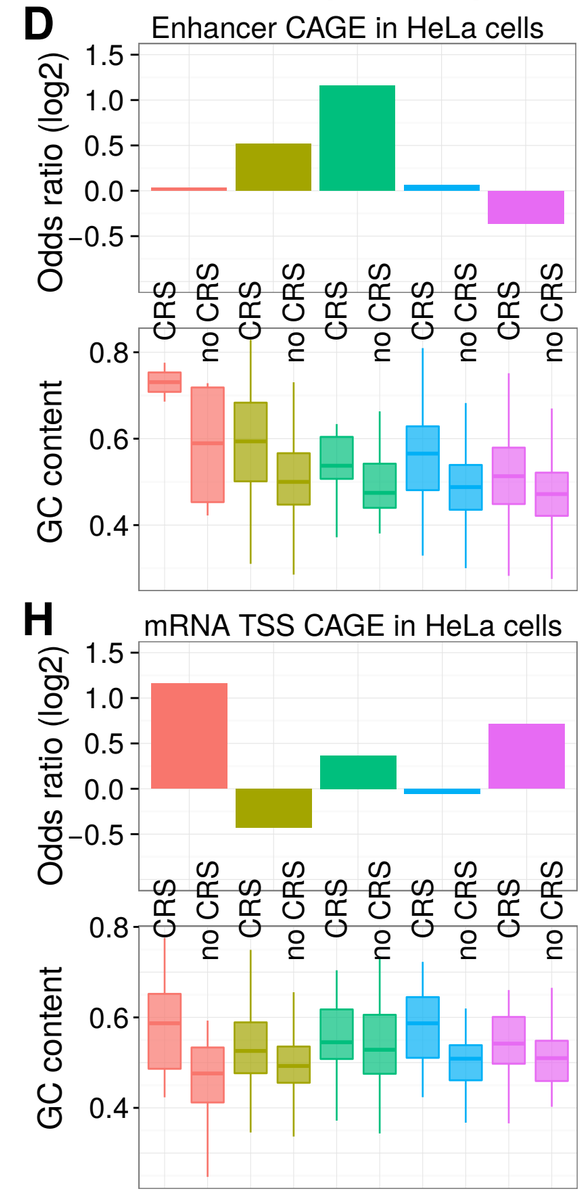
distal TSS of mRNA/lncRNA

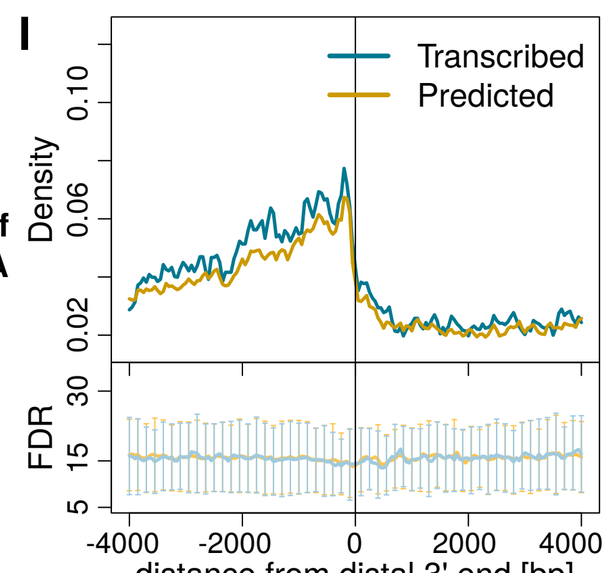
distal 3' end of mRNA/lncRNA

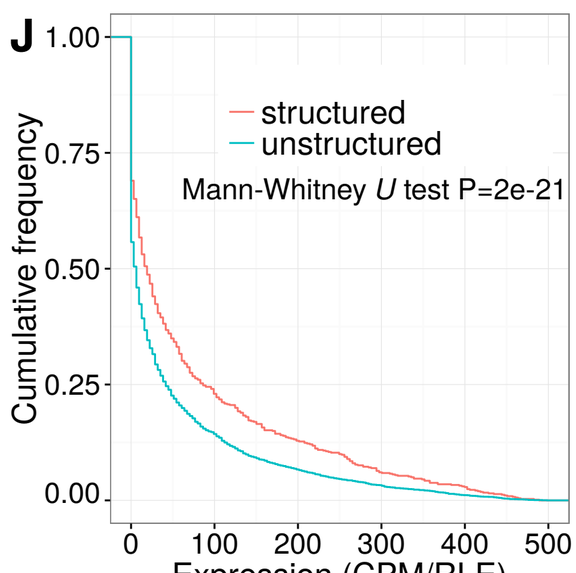


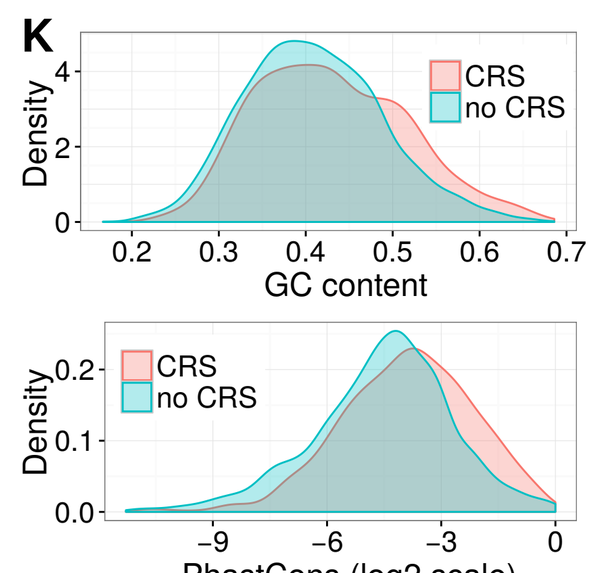






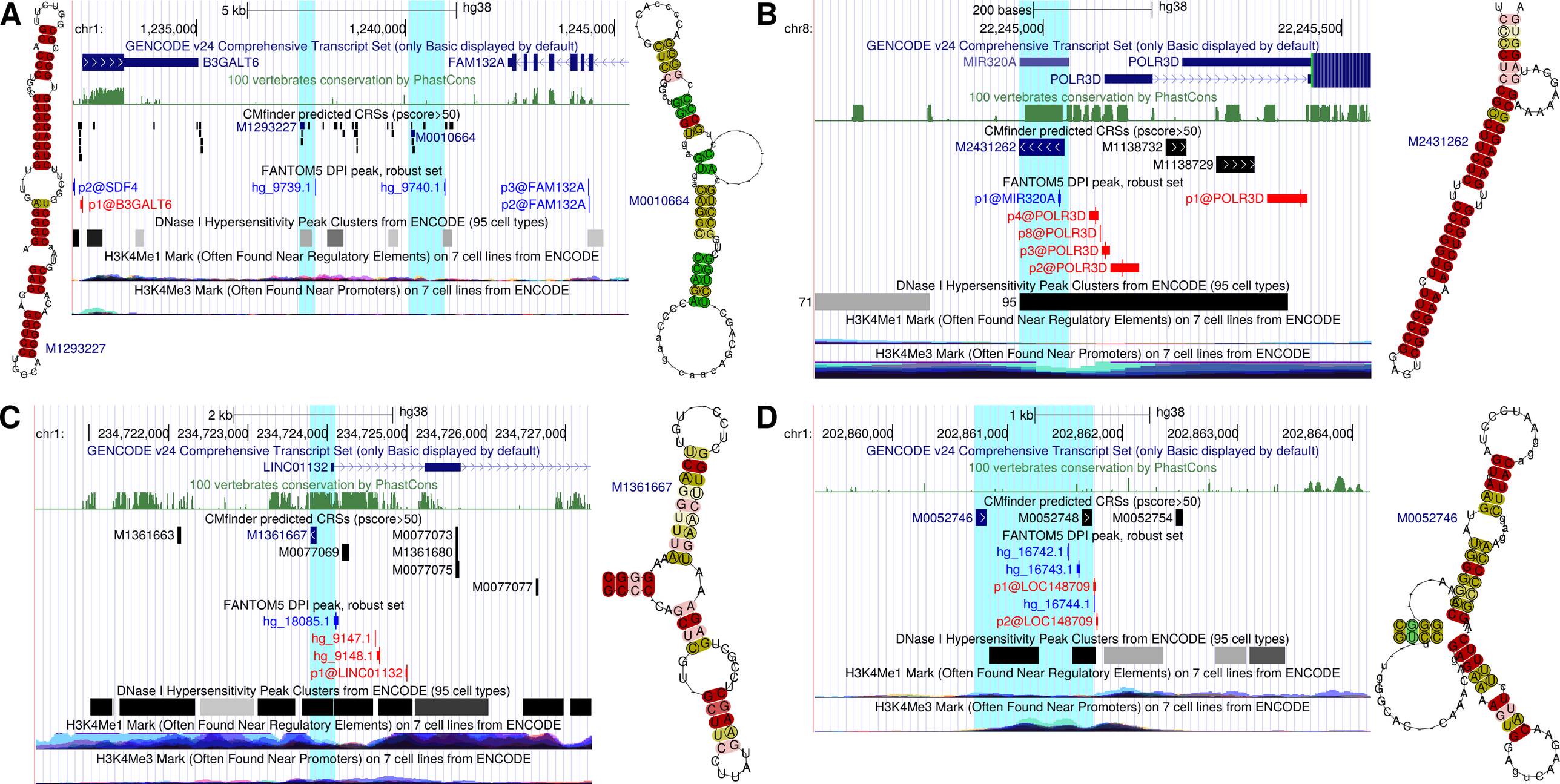






Legend of D and H:

- Bidirectional stable
- Unidirectional stable
- Bidirectional moderate stable
- Unidirectional moderate stable
- Highly unstable


A
5 kb
hg38
chr1:
1,235,000
1,240,000
1,245,000
GENCODE v24 Comprehensive Transcript Set (only Basic displayed by default)
B3GALT6
FAM132A
100 vertebrates conservation by PhastCons
CMfinder predicted CRSs (pscore>50)
M1293227
M0010664
FANTOM5 DPI peak, robust set
p2@SDF4
p1@B3GALT6
hg_9739.1
hg_9740.1
p3@FAM132A
p2@FAM132A
DNase I Hypersensitivity Peak Clusters from ENCODE (95 cell types)
H3K4Me1 Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
H3K4Me3 Mark (Often Found Near Promoters) on 7 cell lines from ENCODE
M1293227
M0010664
B
200 bases
hg38
chr8:
22,245,000
22,245,500
GENCODE v24 Comprehensive Transcript Set (only Basic displayed by default)
MIR320A
POLR3D
POLR3D
100 vertebrates conservation by PhastCons
CMfinder predicted CRSs (pscore>50)
M2431262
M1138732
M1138729
FANTOM5 DPI peak, robust set
p1@MIR320A
p1@POLR3D
p4@POLR3D
p8@POLR3D
p3@POLR3D
p2@POLR3D
DNase I Hypersensitivity Peak Clusters from ENCODE (95 cell types)
71
95
H3K4Me1 Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
H3K4Me3 Mark (Often Found Near Promoters) on 7 cell lines from ENCODE
M2431262
C
2 kb
hg38
chr1:
234,722,000
234,723,000
234,724,000
234,725,000
234,726,000
234,727,000
GENCODE v24 Comprehensive Transcript Set (only Basic displayed by default)
LINC01132
100 vertebrates conservation by PhastCons
CMfinder predicted CRSs (pscore>50)
M1361663
M1361667
M0077069
M0077073
M1361680
M0077075
M0077077
FANTOM5 DPI peak, robust set
hg_18085.1
hg_9147.1
hg_9148.1
p1@LINC01132
DNase I Hypersensitivity Peak Clusters from ENCODE (95 cell types)
H3K4Me1 Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
H3K4Me3 Mark (Often Found Near Promoters) on 7 cell lines from ENCODE
M1361667
D
1 kb
hg38
chr1:
202,860,000
202,861,000
202,862,000
202,863,000
202,864,000
GENCODE v24 Comprehensive Transcript Set (only Basic displayed by default)
100 vertebrates conservation by PhastCons
CMfinder predicted CRSs (pscore>50)
M0052746
M0052748
M0052754
FANTOM5 DPI peak, robust set
hg_16742.1
hg_16743.1
p1@LOC148709
hg_16744.1
p2@LOC148709
DNase I Hypersensitivity Peak Clusters from ENCODE (95 cell types)
H3K4Me1 Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
H3K4Me3 Mark (Often Found Near Promoters) on 7 cell lines from ENCODE
M0052746

The identification and functional annotation of RNA structures conserved in vertebrates

Stefan E Seemann, Aashiq H Mirza, Claus Hansen, et al.

Genome Res. published online May 9, 2017
Access the most recent version at doi:10.1101/gr.208652.116