The snoGloBe interaction predictor enables a broader study of box C/D snoRNA functions and mechanisms

Gabrielle Deschamps-Francoeur[1], Sonia Couture[2], Sherif Abou-Elela[2], Michelle S. Scott[1,*]

[1] Département de biochimie et de génomique fonctionnelle, Faculté de médecine et des sciences de la santé, Université de Sherbrooke, Sherbrooke, Québec, J1E 4K8, Canada

[2] Département de microbiologie et d'infectiologie, Faculté de médecine et des sciences de la santé, Université de Sherbrooke, Sherbrooke, Québec, J1E 4K8, Canada

* Corresponding author: michelle.scott@usherbrooke.ca

Keywords: snoRNA, RNA target prediction, RNA-RNA duplex, gradient boosting classifier, RNA-seq, RNA stability, alternative splicing regulation

ABSTRACT (250 words max)

Box C/D small nucleolar RNAs (snoRNAs) are a conserved class of noncoding RNA known to serve as guides for the site-specific 2'-O-ribose methylation of ribosomal RNAs and the U6 small nuclear RNA, through direct base pairing with the target. In recent years however, several examples of box C/D snoRNAs regulating different levels of gene expression including transcript stability and splicing have been reported. These regulatory interactions typically require direct binding of the target but do not always involve the guide region. Supporting these new box C/D snoRNA functions, high-throughput RNA-RNA interaction datasets detect many interactions between box C/D snoRNAs and messenger RNAs.

To facilitate the study of box C/D snoRNA functionality, we created snoGloBe, a box C/D snoRNA machine learning target predictor based on a gradient boosting classifier and considering snoRNA and target sequence and position as well as target type. SnoGloBe convincingly outperforms general RNA duplex predictors and PLEXY, the only box C/D snoRNA-specific target predictor available. The study of snoGloBe human transcriptome-wide predictions identifies enrichment in snoRNA interactions in exons and on exon-intron junctions. Some specific snoRNAs are predicted to target groups of functionally-related transcripts on common regulatory elements and the exact position of the predicted targets strongly overlaps binding sites of RNA-binding proteins involved in relevant molecular functions. SnoGloBe was also applied to predicting interactions between human box C/D snoRNAs and the SARS-CoV-2 transcriptome, identifying known and novel interactions. Overall, snoGloBe is a timely new tool that will accelerate our understanding of C/D snoRNA targets and function.

INTRODUCTION

Small nucleolar RNAs (snoRNAs) are a conserved class of noncoding RNA present in all eukaryotes. SnoRNAs have been extensively characterized as key players in ribosome biogenesis through the processing and modification of ribosomal RNA (rRNA). They are also implicated in spliceosome biogenesis by guiding the modification of small nuclear RNA (snRNA) (Kiss 2001). To carry out these functions, deemed canonical, they assemble in ribonucleoprotein (snoRNP) complexes which provide stability and catalytic activity to the complex. SnoRNAs are split in two families, the box C/D and the box H/ACA, which respectively guide the 2'-O-ribose methylation and pseudouridylation of RNA, identifying their target sites by base pairing (Weinstein and Steitz 1999). The families differ by their conserved motifs, called boxes, their structure and their interacting proteins.

Box C/D snoRNAs usually range between 50-100 nucleotides and are characterized by their conserved motifs: the boxes C/C' (RUGAUGA) and D/D' (CUGA) (Figure 1A). They interact with core binding proteins SNU13, NOP56, NOP58 and the methyltransferase fibrillarin (FBL) to form the C/D snoRNP. Box C/D snoRNAs guide their catalytic partner to the modification site using sequence complementarity to the region upstream of the boxes D and D', called the antisense element (ASE), ranging from 10 to 20 nucleotides (Kiss-László et al. 1996; Chen et al. 2007). Even though their canonical targets are rRNA and snRNA, some messenger RNAs (mRNAs) have been shown to be targeted using the same mechanism (Kishore and Stamm 2006; Elliott et al. 2019). Also, some snoRNAs, called orphan snoRNAs, have no known canonical target. The bioinformatics tool PLEXY (Kehr et al. 2011) was developed to predict potential box C/D snoRNA targets interacting with the ASE and to reduce the number of orphan snoRNAs. PLEXY only considers the pairing of RNA with the 20 nucleotides upstream of the boxes D and D' of the snoRNA, corresponding to the ASE, and filters the interactions using the pairing

constraints identified in (Chen et al. 2007). PLEXY thus allows to find both rRNA/snRNA and mRNA targets interacting with the ASE.

In recent years, a wide range of functions have been discovered for box C/D snoRNA, including the regulation of chromatin compaction, of metabolic stress, cholesterol trafficking, alternative splicing and mRNA levels (reviewed in (Dupuis-Sandoval et al. 2015; Falaleeva et al. 2017; Bratkovič et al. 2020)). For some of these novel functions, deemed noncanonical, an interaction was reported between the snoRNA and one or multiple target RNA(s), and they are not limited to the ASE. In fact, as shown in Figure 1B, taken together, the interacting regions enabling these functions cover the whole snoRNA, creating a need for a broader tool to predict snoRNA interactions throughout their whole length. In this paper, we define a canonical interaction as an interaction leading to the methylation of a rRNA or a snRNA, and all others are defined as noncanonical, including interactions leading to the methylation of other types of RNA such as mRNA and transfer RNA (tRNA). In support of the noncanonical interactions of C/D snoRNAs, several have been shown to interact with proteins other than the core C/D binding partners mentioned above, such as splicing and polyadenylation factors (Soeno et al. 2010; Kishore et al. 2010; Huang et al. 2017; Zhong et al. 2015). In addition, the accumulation of some snoRNAs is not affected by the depletion of core snoRNP proteins (Deschamps-Francoeur et al. 2014), and some snoRNAs were found in nuclear fractions devoid of such factors (Falaleeva et al. 2016). These studies hint at the possibility that snoRNAs can interact with specific proteins distinct from the core box C/D binding proteins to accomplish specific functions (reviewed in (Baldini et al. 2021; Bergeron et al. 2020; Bratkovič et al. 2020)).

The study of the involvement of snoRNAs in diseases also provides evidence of noncanonical functions and interactors. Although so far, in most cases, snoRNAs have

only been shown to have a deregulated abundance in many diseases and their implication in pathogenesis is typically poorly understood (reviewed in (Zhang et al. 2019; Deogharia and Majumder 2018; Schaffer 2020; Cavaillé 2017)), the molecular mechanisms and pathways involved have been more extensively studied for a small number of pathologies. One snoRNA group proposed as linked to disease for over two decades is the SNORD115 family involved in the Prader-Willi syndrome (PWS), the leading genetic cause of obesity (Cavaillé et al. 2000). The SNORD115 family is lost in most PWS patients (Hebras et al. 2020). The SNORD115 ASE is complementary to the serotonin receptor 2C gene, HTR2C, near a splice site and multiple A to I editing sites (Kishore and Stamm 2006; Vitali et al. 2005). The absence of SNORD115 is thought to lead to a defective serotonin receptor, with consequences on the disease phenotype, although the molecular mechanism is still under debate (Hebras et al. 2020). Another example is SNORD126, which is deregulated in cancer. Its oncogenic role was studied by knockdown and overexpression. SNORD126 was shown to increase the activation of the PI3K-AKT signaling pathway through the upregulation of FGFR2, but the mechanism causing this is still unknown (Fang et al. 2017).

Another form of evidence supporting the involvement of snoRNAs in noncanonical interactions are high-throughput RNA-RNA interaction identification methodologies, PARIS (Lu et al. 2016), LIGR-seq (Sharma et al. 2016) and SPLASH (Aw et al. 2016), which have been devised to enable the survey of all RNA duplexes in cells, both intra- and intermolecular. They have enabled the detection of known and novel snoRNA interactions. For example, LIGR-seq identified functional interactions between orphan snoRNA SNORD83B and three different mRNAs affecting their stability (Sharma et al. 2016). Our de novo analysis of these datasets shows that snoRNAs can interact with diverse RNA biotypes, and mostly with protein coding genes, rRNA and other snoRNAs

(Figure 2A). However, these large-scale methods generally have a low proportion of intermolecular duplex reads (Schönberger et al. 2018) and they are not focused on snoRNAs, suggesting that there are probably plenty of snoRNA-RNA interactions yet to uncover.

To address this question and systematically investigate snoRNA interactors and ultimately the extent of their cellular functions, we developed snoGloBe, a gradient boosting classifier to predict box C/D snoRNA-RNA interactions that involve any region of the snoRNA, not only the ASE. SnoGloBe was trained and tested in human using known canonical interactions, large-scale snoRNA-RNA interactions and validated noncanonical interactions from the literature. The transcriptome-wide prediction of human C/D snoRNA interactions revealed positional enrichment of interactions in targets and functional enrichment for specific snoRNAs, snoGloBe predictions recovering most known snoRNA-mRNA interactions including one described in the SARS-CoV-2 transcriptome. SnoGloBe represents a vital new tool for the characterization of snoRNA functions and mechanisms.

RESULTS

Experimentally detected examples of snoRNA-RNA interactions cover a wide range of interactor biotypes

In order to create a predictor of snoRNA-RNA interactions, we began by assembling all experimentally detected such interactions. We thus performed a manual curation of the literature to obtain these interactions, to which we added all known canonical interactions as compiled in snoRNABase (Lestrade and Weber 2006). This resulted in 149 non-redundant interactions involving snoRNAs, 16 from the literature and 133 from snoRNABase (Figure 2A). To increase and enrich this set, we turned to the high-throughput RNA-RNA interaction (HTRRI) identification methodologies PARIS (Lu et al.

2016), LIGR-seq (Sharma et al. 2016) and SPLASH (Aw et al. 2016). To standardize the analysis of these HTRRI datasets, they were all reanalyzed using the PARIS bioinformatics protocol (Lu et al. 2018) and then filtered to keep only those with a minimum length and without bulges (Figure S1). The list of all interactions obtained from the literature, snoRNABase and from our de novo analysis of HTRRI datasets is available from Table S1. As shown in Figure 2A, the three different sources of positive snoRNA interactions display very different RNA biotype distributions of the targets. While snoRNABase is the main repository of canonical human snoRNA interactions, our manual curation of the literature revealed articles describing noncanonical snoRNA interactions and thus mainly involves protein_coding targets. In contrast, PARIS, LIGR-seq and SPLASH are methodologies detecting RNA-RNA interactions with less user bias. The HTRRI datasets show a much wider distribution of RNA biotypes in terms of distinct interactions, the group with the largest number of distinct targets being protein_coding RNAs, followed by rRNA, snoRNAs, lncRNAs, tRNAs and snRNAs (Figure 2A). Our extensive compilation of snoRNA-RNA interactions thus shows that snoRNAs can interact with many different RNAs of diverse biotypes, supporting the notion that many snoRNA targets and their biological roles in interacting with them remain to be identified. In order to have a full training/testing set for our predictor, we generated negative examples as a combination of random negatives and matched negatives (Figure 2B). The random negative examples are random sequence pairs from any box C/D snoRNA and any gene, whereas the matched negative examples are random sequences originating from a snoRNA-target gene combination from the positive set (Figure 2C). The positive:negative ratio was chosen to be imbalanced (1:20) to reflect the fact that the proportion of transcriptomic sequences bound by C/D snoRNAs is expected to be much lower than the proportion not bound.

Feature encoding and predictor training

SnoRNA-RNA interactions involving the formation of an RNA duplex, the sequences of the two RNAs needed to be encoded amongst the features presented in input. The duplex length of validated functional snoRNA interactions, considering both canonical and noncanonical interactions (Figure 2A, known canonical and known noncanonical), varies from 10 to 32 base pairs, so we chose to encode windows of 13 nucleotides in both the snoRNA and its interacting RNA as a compromise to represent as many validated snoRNA-RNA interactions as possible while limiting the chance of finding these sequences randomly in the genome. More details of the sequence encoding are available from the Methods. The position of the interacting window in the snoRNA is another feature that is important to consider. As mentioned in the Introduction, canonical interactions between snoRNAs and rRNA or snRNAs involve the region immediately upstream of the boxes D or D'. However, validated noncanonical interactions cover collectively the entire snoRNA (Figure 1B). The region employed in the interaction could be an important characteristic for a subgroup of targets and was thus included as an input feature. Similarly, very diverse types of RNAs can be bound by snoRNAs (Figure 2A), with diverse functional outcomes (Bratkovič et al. 2020; Baldini et al. 2021; Bergeron et al. 2020) and we thus chose to include the target biotype as well as the position in the target (either in an exon and/or an intron and whether the exon is a 5' or 3' UTR when appropriate) as input features. Input features are represented schematically in Figure 2D.

Input features were encoded for all positive and negative snoRNA-RNA pairs in the training/test sets. Since snoRNA-RNA pairs are encoded as 13 nucleotide pairs of windows, an interaction can consist of multiple such pairs of windows. In total, the datasets consist of 1838 such positive windows and 38370 negative windows. The overall proportion of positive, random negatives and matched negatives used to train the

predictor is shown in Figure 2B. The datasets were split in a non-overlapping manner for hyperparameter tuning, model training and model testing in a 1:7:2 relative proportion (Figure 2E). Since there are few examples of known noncanonical interactions (Table S1), they were all kept for the test set. SnoRNA genes are often present in multiple copies in genomes and Homo sapiens is no exception (Bergeron et al. 2021; Deschamps-Francoeur et al. 2020). To ensure redundancy is removed from the datasets, they were built such that members of the same family of snoRNAs (as defined by Rfam (Griffiths-Jones et al. 2003)) do not span more than one set (so for example, no member of the SNORD116 family is included in the test set if other members of the family are present in the training set). More details of the implementation are available in the Methods section.

snoGloBe predicts a wide range of interactions with high accuracy

SnoGloBe is based on a gradient boosting classifier, which is a combination of multiple decision trees. This model differs from a random forest by adding each tree successively to lessen the error of the previous ones instead of creating the trees independently. The predictor was built first by tuning the hyperparameters on 10% of the data, using a random search. The model was then trained on 70% of the data using a 5-fold stratified cross-validation (Figure 2E). The output of the prediction is a value between 0 and 1 representing the probability of interaction. The performance of the model was then extensively evaluated on the test set and compared to PLEXY (Kehr et al. 2011) and general RNA-RNA interaction predictors: RNAup (Lorenz et al. 2011), RNAplex (Lorenz et al. 2011), RIsearch2 (Alkan et al. 2017) and IntaRNA (Mann et al. 2017) using parameter values summarized in Figure S2. SnoGloBe displays an excellent separation of the positives and negatives in the independent test set, providing scores below 0.1 for

the great majority (96%) of negative examples and scores above 0.9 for the majority (63%) of positive examples (Figure S3). In addition, it performs considerably better than the other tools on the independent test set, obtaining the highest area under the ROC and precision-recall curves (Figure 3A,B). PLEXY has the weakest performance, which is expected since it only predicts interactions with the ASE and the test set has interactions with all regions of the snoRNA (Figure S4). Interestingly, snoGloBe performs better than general RNA-RNA interaction prediction tools, hinting that snoGloBe captures information specific to snoRNA interactions, more than simple base-pairing. Using thresholds to obtain a 90% precision with every tool, we determined the number of windows predicted as positives and negatives for each tool. SnoGloBe retrieves the highest number of true positive windows, and the highest proportion of known canonical, known noncanonical interactions and HTRRI (Figure 3C,D, S5). PLEXY retrieves the smallest number of positive windows and most are from known canonical interactions, as expected. Generic RNA-RNA interaction predictors give similar results amongst themselves and retrieve in majority known canonical interactions. Interestingly, although snoGloBe was not trained on any known noncanonical interaction, it performs by far best at predicting them, identifying 81/95 noncanonical windows (Figure 3D). Taken together, these data show that snoGloBe enables the prediction of more and a higher diversity of snoRNA-RNA interactions.

Transcriptome-wide predictions using snoGloBe show enrichment of snoRNAs targeting regulatory regions

The interactions of every expressed human snoRNA were predicted against all protein coding genes in human. As many snoRNA copies in human are not expressed and likely represent ‘dead’ copies in the genome, we restricted our study to only those detected as expressed as described in the Methods section. To limit the number of predicted

interactions, we used a stringent cut-off of 3 consecutive windows having a probability (ie snoGloBe output score) greater or equal to 0.98 (Figure S6). With this threshold, we obtain a median of 1017 predicted interactions per snoRNA (Figure 4A). Among those with the largest number of predicted interactions are included several snoRNAs already known to interact functionally with mRNAs such as SNORD32A (>1500 interactions), SNORD83B (>6000 interactions), SNORD88C (>10 000 interactions) (Elliott et al. 2019; Sharma et al. 2016; Scott et al. 2012). The global analysis of all snoRNA predicted interactions reveals a preference for the region upstream of the box D, even though interactions were predicted throughout the whole snoRNA length and the training and test sets rather displayed interactions enriched upstream of both the boxes D and D' (Figure 4B, Figure S4). Interestingly however, individual snoRNAs show different accumulation profiles along their length including some with a clear preference for targets binding the region upstream of the box D' or D and others with a square accumulation in regions other than those upstream of boxes D/D'. For example, SNORD45C in Figure S7 panel A displays two strong regions of target binding, found respectively upstream of the boxes D' and D while SNORD11 in panel B has only one such region, upstream of the box D. In contrast, SNORD31B in panel C and SNORD18A in panel D both have only one clear target binding region, overlapping and downstream of the box D'.

The study of the position of binding sites in targets for individual snoRNAs also reveals interesting trends. Predicted snoRNA interactions are mostly located in introns of their targets, but the proportion of interactions in exons and intron-exon junctions is higher than the transcriptomic composition (Fig 4 compare panels C and D). Indeed, while 7.26% and 0.03% of the coding transcriptome consists of exonic sequences and intron-exon junctions respectively (measured in terms of 13 nt windows as described above), these values go up to 15.65% and 1.19% respectively when only regions predicted to

interact with snoRNAs are considered, an increase of >2 and 40 fold respectively. A high proportion of the exons targeted by snoRNAs are 3' UTR (Figure 4F), which is interesting since some snoRNAs are known to produce microRNAs targeting 3'UTRs (Yu et al. 2015; Lemus-Diaz et al. 2020). The exons targeted by snoRNAs are enriched in 5'UTR and could lead to a role in the regulation of translation initiation (compare Figure E to F). And in general, while exonic regions generally display an increase in the profile of snoRNA predicted interactions per nucleotide (Figure 4G), specific snoRNAs strongly favor binding to precise regulatory regions. For example, SNORD35A and snoU2-30 show strong enrichments respectively on the 3' splice site and the 5' splice site of their targets while SNORD38A shows an enrichment on the polypyrimidine tract (PPT) of its targets (Figure S8).

Some snoRNAs display even more convincing target sets, including strong enrichment in specific gene elements of target genes enriched in specific biological processes, as well as significant overlap with functionally relevant RNA binding protein (RBP) target sites. For example, SNORD50B, which is known to bind K-Ras and regulate its prenylation, was found to be strongly enriched in 5'UTR binding of its targets, using its box D adjacent guide region (Figure S9A,B). Gene ontology enrichment analyses show that predicted targets of SNORD50B are involved in neuronal functionality and genes coding for proteins related to cell-cell interactions (Figure S9C). Many SNORD50B exonic targets bind alternative 5'UTRs, involving alternative transcription start sites, such as those of NDFIP2, COPS3 and SPG21 (Figure S10). In contrast, SNORD22 shows enriched binding to targets on 3' splice sites and PPTs, involving a non-guide region of the snoRNA overlapping the box C' (Figure S11A,B). Gene ontology terms for these targets are enriched in membrane proteins, cell junctions and GTPases (Figure S11C). Many SNORD22 3' splice site targets bind alternatively spliced exons including in the diacylglycerol kinase zeta gene DGKZ, the amyloid beta precursor protein binding family

B member APBB1 and three hits on the same alternatively spliced exon in the focal adhesion protein PXN (Figure S12). Interestingly, when SNORD50B and SNORD22 transcriptome predicted binding sites are compared to RBP binding sites as measured by ENCODE using eCLIPs (Van Nostrand et al. 2020; Davis et al. 2018), they both show strong overlap enrichment with the binding sites of specific RBPs (Figure S9D, S11D). In the case of SNORD50B, the strongest enrichments include DDX3X a helicase known amongst others to bind RNA G-quadruplexes in 5'UTRs, NCBP2 a cap-binding protein interacting with pre-mRNA, BUD13 involved in pre-mRNA splicing and FTO an RNA-demethylase involved in the maturation of mRNAs, tRNAs and snRNAs, supporting a role for SNORD50B in the maturation of pre-mRNA and in particular their 5' extremity. In contrast, for SNORD22, the RBPs with strongest enrichment are PCBP2 the poly(rC) binding protein, PTBP1 a PPT binding protein involved in the regulation of alternative splicing as well as BUD13, PRPF8 and AQR all known as involved in pre-mRNA splicing, supporting a role for SNORD22 in the regulation of alternative splicing, in particular through the binding of PPTs.

SnoGloBe predicts SNORD126 interactions that can be validated experimentally

To further investigate snoGloBe's predictions, we focused on SNORD126. SNORD126 is listed as an orphan snoRNA in snoRNABase (Lestrade and Weber 2006), but is predicted to interact with 28S rRNA in snoRNA Atlas (Jorjani et al. 2016) and was shown to methylate this target in rats in a tissue-specific manner (Hebras et al. 2019). SNORD126 was shown to activate the PI3K-AKT pathway, conferring it an oncogenic role, though the underlying molecular mechanism is still unknown (Fang et al. 2017, 126). The SNORD126 interactions predicted by snoGloBe against protein coding genes led to interesting profiles. The most represented interaction region in SNORD126 is

located in the middle of the snoRNA, overlapping the D' box, which doesn't correspond to an ASE (Figure 5A), suggesting a molecular mechanism of target binding different from the canonical mechanism. The predicted folding of SNORD126 forcing the predicted interacting region to remain single stranded is shown in Figure 5B. SNORD126 interactions are enriched in exons (1.6 fold) and particularly enriched on intron-exon junctions (>50 fold) compared to the protein coding transcriptome composition (Figure 5C-D). The exons predicted to be targeted by SNORD126 are enriched in 5'UTR (Figure 5E-F). SNORD126 predicted interactions seem somewhat uniformly distributed in the target exons, but are enriched around 80 nucleotides upstream of the exons (Figure 5G). Taken together, these data suggest that SNORD126 could play a role in the regulation of RNA stability and splicing.

To investigate this hypothesis, we studied the effect of SNORD126 knockdown on the predicted target genes by RNA-seq. Indeed, we found that the knockdown of SNORD126 significantly affects the RNA level of 1050 protein coding genes including 65 predicted targets (Fig. 5H), showing a significant enrichment (p-value < 0.0001 by random sampling analysis as described in the Methods). The genes affected by the knockdown are mostly downregulated (Fig. 5I). As for the regulation of alternative splicing, 309 such events are affected by the knockdown of SNORD126, 25 of which overlap a predicted SNORD126 binding site (p-value = 0.0003). Amongst the interesting candidates, the target site of SNORD126 on CPT1B overlaps an alternative 5' splice site detected with a differential splicing pattern and the target site of SNORD126 on MR1 is near a 3' splice site for which the intron has an alternative 5' extremity (Figure S13). Interestingly, three genes having an alternative splicing event overlapping a predicted interaction were also differentially expressed upon SNORD126 knockdown: CPT1B, MR1 and DDX11. In contrast to snoGloBe predictions of SNORD126 for which a total of 72 were detected as either stability and/or splicing targets following SNORD126

knockdown, only two SNORD126 interactions were identified in the HTRRI datasets, one with a mRNA and one with the 18S rRNA, neither of which is detected as affected by the SNORD126 knockdown, emphasizing the usefulness of snoGloBe in addition to high-throughput methodologies.

SnoGloBe predicts interactions between human snoRNAs and the SARS-CoV-2 transcriptome

As another example of the utility of snoGloBe, we applied it to the SARS-CoV-2 transcriptome. It has been shown that the SARS-CoV-2 genome is heavily 2'-O-ribose methylated and interacts strongly with snoRNAs using the high-throughput structure probing methodology SPLASH (Aw et al. 2016; Yang et al. 2021). Thus as another proof of concept for the utility and capacity of snoGloBe, we predicted the interactions between human snoRNAs and the SARS-CoV-2 genome using snoGloBe. We detected 8818 interactions between 312 snoRNAs and the SARS-CoV-2 genome, the distribution of which is shown in Figure 6A. One of the strongest interaction partners between SARS-CoV-2 and host transcripts is reported to be with the box C/D snoRNA SNORD27 (Yang et al. 2021). Although, the SPLASH experiments were carried out in Vero-E6 cells from African green monkey kidney, snoGloBe detects the SARS-CoV-2 interaction with human SNORD27 (Figure 6B), suggesting that the interaction is also relevant in human.

SnoGloBe availability and usage

The snoGloBe code is written in Python using the machine learning package scikit-learn (Pedregosa et al. 2011). It is freely available and can be downloaded from gitlabscottgroup.med.usherbrooke.ca/scott-group/snoglobe. Users must provide a file with the sequences of the snoRNAs of interest, the sequences of whole chromosomes,

an annotation file in gtf format and a file with the potential target identifiers to scan for snoRNA interactions. Detailed instructions are available in the help manual.

DISCUSSION

Motivated by the continually increasing number of examples of snoRNAs interacting with noncanonical targets using diverse regions within but also without the ASE (Figure 1) as well as by the diversity in RNAs targeted by snoRNAs (Figure 2A), we built snoGloBe, a box C/D snoRNA interaction predictor that considers the whole snoRNA and any type of RNA target. SnoGloBe is a gradient boosting classifier that takes into account the sequence of the snoRNA and its potential target as well as the position in the snoRNA and the type and position in the potential target. Compared to general use RNA-RNA interaction predictors that consider only sequence complementarity and the interaction stability of the duplex, snoGloBe performs considerably better, suggesting that considering snoRNA and target features enhances the prediction. SnoGloBe also performs better than the snoRNA-specific predictor PLEXY which was not built to predict interactions outside of the snoRNA ASEs. Many such non-ASE interactions are detected in HTRRI datasets and some have been extensively validated for individual snoRNAs (Figure 1B), limiting the scope of the PLEXY predictor. Interestingly a subset of positive examples (30%) is not found by snoGloBe and while this proportion is considerably lower than for all other predictors considered (they miss >70% of positives in the test set, Figure 3D), there is still room for improvement. This subset involves mostly interactions displaying bulges in the base pairing in one or both members of the interaction, which is more difficult to accurately identify and will require different approaches and likely larger training datasets for machine learning approaches to accurately predict them.

The study of the snoGloBe predicted interactions in human is very interesting and opens numerous research avenues that will likely lead to important insights into snoRNA function. Dozens of snoRNAs display profiles supporting the non-uniform distribution of predicted targets in pre-mRNA, with hundreds or even thousands of targets enriched in common regulatory elements such as PPTs, 5' or 3' splice sites and 5' or 3' UTRs (Figures 4, 5, S8, S9, S11). Each such snoRNA target profile will require in depth integrative analysis to consider the possible functionality, molecular mechanism and ultimately cellular outcome of the collective regulation of these targets by the snoRNA. We began such studies for SNORD50B, SNORD22 and SNORD126, all three of which display strong enrichment for binding to specific regulatory elements, respectively 5' UTRs, PPTs/3' splice sites and 3' of introns. Manual review of their predicted targets led to the identification of a subset of such binding events overlapping alternatively regulated events (for example alternative 5' UTRs for SNORD50B and alternatively spliced exons for SNORD22, Figures S10, S12). Gene ontology analysis of the targets show strong enrichment for specific biological processes and provide convincing subsets of targets to focus on. Finally the strong overlap between snoRNA predicted binding sites and ENCODE-detected RBP binding sites is important evidence of the functional relevance of the interactions, particularly as the function of the RBPs with the strongest overlap strongly supports the type of regulation likely carried out by the snoRNA. Several molecular mechanisms could explain the binding overlap of snoRNA and RBP at the same position on the same target pre-mRNA. The snoRNA could be guiding the RBP to its target as snoRNAs do for core snoRNA binding proteins such as FBL, and as has been shown for nuclear exosome components as shown in (Zhong et al. 2015). However, since RNA binding motifs are known for several of the RBP considered and because snoRNAs have not been found as enriched binding partners of all these RBPs, it is likely that the snoRNAs and some of the RBPs are competing for the same binding

site. Further studies will be required to define the snoRNA-RBP relationship and its effect on the regulation of the targets. These overlapping snoRNA-RBP targets could be revealing novel levels of post-transcriptional regulation, the understanding of which will be important in health and disease.

Overall, while HTRRI datasets have collectively been generated in a handful of cell lines, considerable time and money would be required to explore normal human tissues and diverse conditions using these methodologies. SnoGloBe predictions will be instrumental in filling the gap by providing rapid predictions for snoRNA interactions that can then be further investigated to better understand cellular functionality. Our additional demonstration that snoGloBe can be used to investigate the interactions between snoRNAs and viral transcripts (Figure 6), further widens its scope and utility.

METHODS

High-throughput RNA-RNA interaction analysis

The high-throughput RNA-RNA interaction datasets from PARIS (SRR2814761, SRR2814762, SRR2814763, SRR2814764 and SRR2814765), LIGR-seq (SRR3361013 and SRR3361017) and SPLASH (SRR3404924, SRR3404925, SRR3404936 and SRR3404937) were obtained from the short read archive SRA (https://www.ncbi.nlm.nih.gov/sra) using fastq-dump from the SRA toolkit (v2.8.2). The PARIS datasets were trimmed using the icSHAPE pipeline available at https://github.com/qczhang/icSHAPE. PCR duplicates were removed from LIGR-seq datasets using the script readCollpase.pl from the icSHAPE pipeline and the reads were trimmed using Trimmomatic version 0.35 with the following options : HEADCROP:5 ILLUMINACLIP:TruSeq3-SE.fa:2:30:4 TRAILING:20 MINLEN:25. The quality of the reads was assessed using FastQC (v0.11.15) before and after the pre-processing steps.

All the samples were analyzed using the PARIS pipeline as described in sections 3.7 and 3.8 from (Lu et al. 2018). Some modifications were made to the duplex identification and annotation scripts. The modified scripts are available at gitlabscottgroup.med.usherbrooke.ca/gabrielle/paris_pipeline.

The RNA duplexes were assigned to genes using the annotation file described in (Boivin et al. 2018) to which missing rRNA annotations from RefSeq were added. The annotation file was modified using CoCo correct_annotation (Deschamps-Francoeur et al. 2019) to ensure the correct identification of snoRNA interactions.

Only the interactions between two known genes including a box C/D snoRNA were kept. To avoid intramolecular interactions, we removed interactions between a snoRNA and its 50 flanking nucleotides and interactions between two snoRNAs of a same Rfam family (Griffiths-Jones et al. 2003). To limit the number of false positives, we filtered the interactions based on their pairing using RNAplex (Lorenz et al. 2011). Only paired regions of the interactions were kept to get rid of unpaired flanking regions. The interactions were split at each bulge to ensure the correct alignment of the snoRNA and target sequences. The interactions shorter than 13 nucleotides were removed to respect the length of the windows (see Input features section). We finally removed interactions that were already known and present in our positive set described in the next section. We obtained 445 box C/D snoRNA interactions (Fig S1).

Positive set composition

The positive set is composed of the previously detected snoRNA interactions from PARIS, LIGR-seq and SPLASH filtered as described above, as well as interactions obtained from snoRNABase (Lestrade and Weber 2006) and manually curated interactions from the literature (Table S1) (Fig 2B). Interactions from snoRNABase and

from the literature that were shorter than 13 nucleotides were padded by adding their flanking sequence to respect the length threshold.

Negative set composition

The negative set is composed of random negatives and matched negatives (Fig 2B). The random negative examples are the combination of random sequences from any box C/D snoRNA and any gene, whereas the matched negative examples are random sequences coming from a positive snoRNA-target gene combination (Fig 2C).

Input features

The interactions were split in 13 nucleotides sliding windows, with a step of 1 nucleotide. The 13 nucleotide window length was chosen to limit the chance of finding this sequence randomly in the genome, and most of the known interactions respect this length. Each interaction window is composed of 13 nucleotides of the snoRNA and the corresponding 13 nucleotides of the target. The input features used are the window sequences in one-hot encoding, the relative position in the snoRNA between 0 and 1, the location in the target gene (intron, exon, 3'UTR and/or 5'UTR) and the target biotype (Fig 2D). The biotypes considered are listed in snoGloBe's manual. Protein coding, pseudogene and long noncoding RNA biotypes were grouped according to http://ensembl.org/Help/Faq?id=468 and http://vega.archive.ensembl.org/info/about/gene_and_transcript_types.html.

Redundancy removal for tuning, training and test sets

The positive and negative examples were split into hyperparameter tuning, training and test sets. First, to remove redundancy from the sets, the snoRNAs were grouped based on their Rfam identifier (Griffiths-Jones et al. 2003). To ensure that the model is not trained and tested on similar snoRNAs, the Rfam families were split in order to assign

20% of all examples in the test sets, with similar proportion of the initial high-throughput RNA-RNA interactions and known canonical interactions. All the known noncanonical interactions were kept for the test sets since there are very few such examples. The remaining examples were split between the hyperparameter tuning and training set to get respectively 10% and 70% of initial data. (Fig 2E)

Building the model

The model used is a gradient boosting classifier from scikit-learn (v0.21.3) (Pedregosa et al. 2011). We selected the hyperparameters using a random search with 3-fold cross-validation. The selected hyperparameters are: n_estimators = 371, min_samples_split = 76, min_samples_leaf = 49, max_depth = 2 and learning_rate = 0.43, others are kept to default.

The model was trained on the whole training set and validated with 5-fold stratified cross-validation, to keep similar proportions of interactions coming from high-throughput methodologies and from snoRNABase in each subset. The resulting model is called snoGloBe.

The model performance was evaluated on the test set and compared to PLEXY (Kehr et al. 2011), RNAplex (Lorenz et al. 2011), RIsearch2 (Alkan et al. 2017), IntaRNA (Mann et al. 2017) and RNAup (Lorenz et al. 2011). To compare their performance on a similar basis, we selected a threshold (either a score for snoGloBe, or an energetic cut-off for the other tools) resulting in 90% precision on the test set. The details are available in Figure S2. PLEXY was only used for snoRNAs with non-degenerated boxes D and D' to avoid bias caused by misidentified boxes.

Prediction against protein coding genes

SnoGloBe was used to predict box C/D snoRNA interactions with protein coding transcripts. For this analysis, only box C/D snoRNAs expressed at 1 transcript per million (TPM) or more in at least one of the RNA-seq datasets from 7 different healthy human tissues (3 samples from different individuals for each of the following tissues: brain, breast, liver, ovary, prostate, skeletal muscle and testis) from a previous study (Fafard-Couture et al. 2021) (available from GEO: GSE126797, GSE157846) were considered. We predicted the interactions of these snoRNAs against all protein coding regions of the genome, split in 13-nucleotide windows with a step of two. We took whole gene sequences to predict interactions with any intron and exon. To narrow the number of predictions obtained, we kept the interactions having at least three consecutive windows with a score >= 0.98 for further analysis. The gene ontology enrichment analysis of the predicted targets was done using g:Profiler (Raudvere et al. 2019).

Overlap between predicted snoRNA interactions and eCLIP region

All the eCLIP datasets (Van Nostrand et al. 2020, 2016) were downloaded from the ENCODE portal (Davis et al. 2018), totaling 225 samples of 150 proteins. The complete list of the datasets is available in Table S2. Only the eCLIP regions having a p-value <= 0.01 were kept. Datasets from the same protein were merged using BEDTools merge -s (v2.26.0) (Quinlan and Hall 2010). The number of overlaps between the predicted interactions and the eCLIP regions was computed using BEDTools intersect -s.

SNORD126 knockdown

HepG2 cells were cultured in complete Eagle's Minimum Essential Medium (EMEM from Wisent) and passaged twice a week, according to ATCC guidelines. Trypsinized cells were then seeded at 350000 cells/well in 6 well plates in 1ml EMEM. Cells were

transfected 24 hours later with 2 different ASOs targeting SNORD126 (30nM or 40nM) using Lipofectamine 2000 (LIFE technologies) and optiMEM (Wisent). A scrambled ASO was used as a negative control. The sequence of the ASOs are listed in Supplemental Figure S14.

Cells were harvested 48 hours post transfection, washed and pelleted, then resuspended in 1ml Trizol and stored at -80℃ until RNA extraction. This was repeated 3 times to obtain biological triplicates.

RNA extraction

Total RNA extraction from transfected HepG2 cells was performed using RNeasy mini kit (Qiagen) as recommended by the manufacturer including on column DNase digestion with RNase-Free DNase Set (Qiagen). However, 1.5 volumes Ethanol 100% was used instead of the recommended 1 volume ethanol 70% in order to retain smaller RNA. RNA integrity of each sample was assessed with an Agilent 2100 Bioanalyzer.

RNA was reversed transcribed using Transcriptor reverse transcriptase (Roche) and knockdown levels were evaluated by qPCR.

RNA-seq library preparation and sequencing

RNAseq librairies were generated from 1ug DNA-free total RNA/condition using the NEBNext® Ultra™ II Directional RNA Library Prep Kit for Illumina (E7760S) and following the Protocol for use with NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB #E7490). The resulting librairies were submitted to a total of 10 cycles of amplification then purified using 0.9X Ampure XP beads. Quality and size was assessed with an Agilent 2100 Bioanalyser. Librairies were then quantified using a Qubit fluorometer, pooled at equimolar concentration and 1.8pM was sequenced on Illumina's

NextSeq 500 using a NextSeq 500/550 High Output Kit v2.5 (150 cycles) paired-end 2x75bp.

RNA-seq analysis

The resulting base calls were converted to fastq files using bcl2fastq v2.20 (Illumina) with the following options: --minimum-trimmed-read-length 13, --mask-short-adapter-reads 13, --no-lane-splitting. The fastq files were trimmed for quality and to remove remaining adapters using Trimmomatic v0.36 (Bolger et al. 2014) with ILLUMINACLIP: {adapter_fasta}:2:12:10:8:true, TRAILING:30, LEADING:30, MINLEN:3. The sequence quality was assessed using FastQC v0.11.5 (Andrews) before and after the trimming. The trimmed sequences were aligned to the human genome (hg38) using STAR v2.6.1a (Dobin et al. 2013) with the options --outFilterScoreMinOverLread 0.3, --outFilterMatchNminOverLread 0.3, --outFilterMultimapNmax 100, --winAnchorMultimapNmax 10, --alignEndsProtrude 5 ConcordantPair. Only primary alignments were kept using samtools view -F 256 (v1.5) (Li et al. 2009). The gene quantification was done using CoCo correct_count -s 2 -p (v0.2.5p1) (Deschamps-Francoeur et al. 2019). DESeq2 (Love et al. 2014) was used for the differential expression analysis. Genes having a corrected p-value <= 0.01 were considered significantly differentially expressed. The alternative splicing analysis was done using MAJIQ v2.2 and VOILA with the option --threshold 0.1 (Vaquero-Garcia et al. 2016).

Empirical p-value calculation

To evaluate the significance of the overlaps between each snoRNA predicted interactions and eCLIP binding sites, alternative splicing events and differentially expressed genes, we computed an empirical p-value using BEDTools shuffle 10 to

100 000 times for each combination followed by BEDTools intersect -s through pybedtools (Quinlan and Hall 2010; Dale et al. 2011). BEDTools shuffle was used with an appropriate background for each analysis: all protein coding genes for eCLIP binding sites and protein coding genes having an average of 1TPM across all sequencing datasets for differential expression and alternative splicing analyses. For eCLIP binding sites and alternative splicing events, we counted the number of overlaps between the shuffled experiments and the region of the events, whereas for differential expressed genes, we counted the number of overlaps between the shuffled experiments and the entire differentially expressed genes with no regards on the position of the overlap in the gene. P-values were calculated as the proportion of iterations in which the shuffled dataset overlap was at least as extreme as the true dataset overlap.

Prediction of human snoRNA interaction with SARS-CoV-2 transcriptome

We predicted the interaction between the expressed human snoRNAs against SARS-CoV-2 genes, using SARS-CoV-2 ASM985889v3 genome assembly and the annotation file Sars_cov_2.ASM985889v3.101.gtf obtained from Ensembl COVID-19. We used thresholds of minimum 3 consecutive windows having a probability greater or equal to 0.85.

COMPETING INTERESTS STATEMENT

The authors declare no competing interests.

ACKNOWLEDGEMENTS

The authors would like to thank members of the Scott and Abou-Elela groups for helpful discussions and Compute Canada for providing state-of-the-art computing infrastructures. This work was supported by a CIHR grant (PJT 153171) to MSS and

SAE and a team FRQ-NT grant to MSS and SAE. GDF was supported by a NSERC Doctoral scholarship. MSS holds a Fonds de Recherche du Québec – Santé (FRQ-S) Research Scholar Senior Career Award.

Figure legends

Figure 1. Box C/D snoRNA characteristics and interactions. A) Box C/D snoRNAs have well conserved patterns, called boxes C/C' (green) and D/D' (purple). The canonical interaction with the target (orange) occurs upstream of the boxes D and D', using a region referred to as the ASE. The fifth nucleotide upstream of these boxes is methylated by the core C/D snoRNA interactor FBL. B) Schematic representation of experimentally validated noncanonical interaction regions. The classical elements are represented: the boxes C and C' (green), D and D' (purple) and the ASE (yellow). Bipartite interactions are represented by a dotted line. The color of the interaction represents its effect, whether regulation of alternative splicing, regulation of RNA level (either pre-mRNA or mRNA) and RNA modification including methylation of noncanonical targets, acetylation and A to I editing. Combined together, noncanonical interactions cover the whole snoRNA.

Figure 2. Composition of the dataset used to build snoGloBe. A) Diverse RNAs have been shown to bind box C/D snoRNAs. Interactions involving box C/D snoRNA were collected and assembled including known canonical interactions with rRNA and snRNA from snoRNABase, known noncanonical interactions curated from the literature and interactions extracted from HTRRI datasets. The proportion of interactions involving different RNAs of each biotype is shown for each interaction source. The color legend for RNA biotypes is shown on the right. (B) Distribution of the datasets used to build snoGloBe. The dataset consists of positive, matched negatives and random negatives in a proportion of 20 negatives for 1 positive window. The positive windows are composed of HTRRI (86.3 %), known canonical (8.5 %) and noncanonical (5.2 %) interactions. (C) Generation of matched negative windows. 10 matched negative windows are generated for each positive one. The matched negative windows originate from the same snoRNA–

target gene pair as the positive window. One has the same position in the snoRNA and a different position in the same target, one has a different position in the snoRNA and the same position in the target, and 8 windows have random positions in the same snoRNA-target pair. D) SnoRNA-RNA pairs are encoded for presentation to the predictor. Features considered include the 13 nt sequence of the snoRNA and the 13 nt sequence of the target, the relative position of the window in the snoRNA, the target biotype and the position in the target. E) The dataset is split in non-overlapping sets for hyperparameter tuning set (10% of the windows), training set (70% of the windows) and test set (20% of the windows). The hyperparameter tuning was done using a random search with 3-fold cross-validation. The model was trained and evaluated using stratified 5-fold cross-validation to ensure the correct representation of each category of positive windows in each subset. The known noncanonical windows were all kept for the validation set.

Figure 3. snoGloBe performs better than the other tools tested on the test set. (A) Receiver Operator Characteristic (ROC) and (B) Precision-Recall (PR) curves of different tools calculated on the test set. The corresponding area under the curves (AUC) are indicated in the legend. C) Table of performance measures from different tools calculated on the test set with a threshold set to obtain a precision of 90%. PLEXY was only used on interactions from box C/D snoRNA with non-degenerated boxes D and D' (Table S1) since the position and sequence of the boxes are required. D) Upset plot representing the overlaps between each tool prediction of the test set's positive windows. The upset plot only shows the subset of the interactions that were used for PLEXY to ensure a fair comparison. The upset plot of all the test set's positive windows is show in Fig S5.

Figure 4. Box C/D snoRNA predicted interactions across the coding transcriptome. A) Histogram and boxplot (above) of the number of interactions per snoRNA using a threshold of at least 3 consecutive windows having a probability greater or equal to 0.98. Most snoRNA have less than 2000 predicted interactions. B) Distribution of the predicted region of interaction in all snoRNAs. The position in the snoRNAs is normalized between 0 and 1. The computationally identified boxes C, D' and D are respectively represented in green and purple. The predicted interactions are found throughout the snoRNA, with an enrichment in the 3' end. C-D) Bar chart representing the proportion of exon, intron and intron-exon junction in the protein transcriptome (C) and the box C/D predicted interactions in the target (D). The predicted interactions are enriched in the exons and the intron-exon junctions (D) compared to the protein coding transcriptome (C). E-F) Doughnut charts representing the composition in terms of 13-nt windows of the exons of the protein coding transcriptome (E) and the box C/D snoRNA predicted interactions in the targets (F). The predicted interactions located in exons are mainly found in UTRs (F) and are enriched in 5'UTRs when compared to the protein coding transcriptome (E). G) Distribution of the predicted interactions 100 nucleotides upstream of exons (left), in the exon (middle), and 100 nucleotides downstream of the exons (right). The positions in the exons are normalized between 0 and 1. The number of interactions is normalized by the number of existing features (exons or introns) at each position. The predicted interactions are uniformly distributed across the exons, there is a higher number of interactions predicted inside the exons than in the flanking nucleotides.

Figure 5. SNORD126 predicted targets are significantly affected by its knockdown. A) The major interaction site in SNORD126 predicted by snoGloBe is located in the middle of the snoRNA and doesn't match the ASE upstream of the boxes D and D'

represented in purple. The accumulation profile represents the proportion of SNORD126 predicted interactions overlapping each nucleotide in the snoRNA. B) Predicted folded structure of SNORD126 considering the main region of interaction. Mfold (Zuker 2003) was used to predict the secondary structure of SNORD126, forcing nucleotides 37 to 53 to be single stranded (blue) through the unafold webserver. C-D) SNORD126 predicted interactions are enriched in the exons and the intron-exon junctions. (C) Shows the relative length of the different elements of the protein coding transcriptome while (D) shows the relative proportion of the targets of SNORD126. E-F) The predicted interactions located in exons are enriched in 5'UTRs. (E) Doughnut chart showing the breakdown of the different constituents of exons in the protein coding transcriptome. (F) Doughnut chart showing the same breakdown but only for regions targeted by SNORD126. G) SNORD126 predicted interactions are uniformly distributed across exons, with an enrichment around 80 nucleotides upstream of exons. H) Venn diagram representing the overlap between significantly differentially expressed genes following SNORD126 knockdown in red and SNORD126 predicted targets (green). The overlap between SNORD126 predicted targets and the differentially expressed genes (orange) is significant, with an empirical p-value < 0.0001. I) Volcano plot representing the impact of SNORD126 knockdown on protein coding genes. The dots above the gray line are considered significantly differentially expressed (adjusted p-value ≤ 0.01). SNORD126 predicted targets are colored in blue. (J) Venn diagram representing the overlap between alternative splicing events following SNORD126 knockdown (blue) and SNORD126 predicted interactions (orange). The predicted interaction must be inside the alternative splicing event to be considered as overlapping. The overlap between SNORD126 predicted interactions and the alternative splicing events is significant with an empirical p-value = 0.0005. Only genes having a mean of 1 TPM across all samples are shown in (H-J).

Figure 6. Human box C/D snoRNAs are predicted to target the SARS-CoV-2 transcriptome. A) Distribution of the number of predicted interactions per snoRNA across the SARS-CoV-2 transcriptome. Human box C/D snoRNAs have a median of 22 predicted interactions having at least 3 consecutive windows with a score greater or equal to 0.85 with SARS-CoV-2 transcriptome. B) The validated interaction between African green monkey SNORD27 and SARS-CoV-2 is also predicted with human SNORD27. The validated interaction is shown in red, the nucleotide that differs between the African green monkey and human SNORD27 is underlined. The predicted interaction is outlined by the box.

REFERENCES

Alkan F, Wenzel A, Palasca O, Kerpedjiev P, Rudebeck AF, Stadler PF, Hofacker IL, Gorodkin J. 2017. RIsearch2: suffix array-based large-scale prediction of RNA-RNA interactions and siRNA off-targets. *Nucleic Acids Res* **45**: e60.

Andrews S. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (Accessed July 5, 2021).

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

Baldini L, Charpentier B, Labialle S. 2021. Emerging Data on the Diversity of Molecular Mechanisms Involving C/D snoRNAs. *Non-Coding RNA* **7**: 30.

Bergeron D, Fafard-Couture É, Scott MS. 2020. Small nucleolar RNAs: continuing identification of novel members and increasing diversity of their molecular mechanisms of action. *Biochem Soc Trans* **48**: 645–656.

Bergeron D, Laforest C, Carpentier S, Calvé A, Fafard-Couture É, Deschamps-Francoeur G, Scott MS. 2021. SnoRNA copy regulation affects family size, genomic location and family abundance levels. *BMC Genomics* **22**: 414.

Boivin V, Deschamps-Francoeur G, Couture S, Nottingham RM, Bouchard-Bourelle P, Lambowitz AM, Scott MS, Abou-Elela S. 2018. Simultaneous sequencing of coding and noncoding RNA reveals a human transcriptome dominated by a small number of highly expressed noncoding genes. *RNA N Y N* **24**: 950–965.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.

Bratkovič T, Božič J, Rogelj B. 2020. Functional diversity of small nucleolar RNAs. *Nucleic Acids Res* **48**: 1627–1651.

Cavaillé J. 2017. Box C/D small nucleolar RNA genes and the Prader-Willi syndrome: a complex interplay. *Wiley Interdiscip Rev RNA* **8**.

Cavaillé J, Buiting K, Kiefmann M, Lalande M, Brannan CI, Horsthemke B, Bachellerie J-P, Brosius J, Hüttenhofer A. 2000. Identification of brain-specific

and imprinted small nucleolar RNA genes exhibiting an unusual genomic organization. *Proc Natl Acad Sci U S A* **97**: 14311–14316.

Chen C-L, Perasso R, Qu L-H, Amar L. 2007. Exploration of Pairing Constraints Identifies a 9 Base-pair Core within Box C/D snoRNA–rRNA Duplexes. *J Mol Biol* **369**: 771–783.

Dale RK, Pedersen BS, Quinlan AR. 2011. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinforma Oxf Engl* **27**: 3423–3424.

Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, Hilton JA, Jain K, Baymuradov UK, Narayanan AK, et al. 2018. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res* **46**: D794–D801.

Deogharia M, Majumder M. 2018. Guide snoRNAs: Drivers or Passengers in Human Disease? *Biology* **8**: E1.

Deschamps-Francoeur G, Boivin V, Abou Elela S, Scott MS. 2019. CoCo: RNA-seq read assignment correction for nested genes and multimapped reads. *Bioinformatics* **35**: 5039–5047.

Deschamps-Francoeur G, Garneau D, Dupuis-Sandoval F, Roy A, Frappier M, Catala M, Couture S, Barbe-Marcoux M, Abou-Elela S, Scott MS. 2014. Identification of discrete classes of small nucleolar RNA featuring different ends and RNA binding protein dependency. *Nucleic Acids Res* **42**: 10073–10085.

Deschamps-Francoeur G, Simoneau J, Scott MS. 2020. Handling multi-mapped reads in RNA-seq. *Comput Struct Biotechnol J* **18**: 1569–1576.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21.

Dupuis-Sandoval F, Poirier M, Scott MS. 2015. The emerging landscape of small nucleolar RNAs in cell biology. *Wiley Interdiscip Rev RNA* **6**: 381–397.

Elliott BA, Ho H-T, Ranganathan SV, Vangaveti S, Ilkayeva O, Assi HA, Choi AK, Agris PF, Holley CL. 2019. Modification of messenger RNA by 2′- O -methylation regulates gene expression in vivo. *Nat Commun* **10**: 1–9.

Fafard-Couture É, Bergeron D, Couture S, Abou-Elela S, Scott MS. 2021. Annotation of snoRNA abundance across human tissues reveals complex snoRNA-host gene relationships. *Genome Biol* **22**: 172.

Falaleeva M, Pages A, Matuszek Z, Hidmi S, Agranat-Tamir L, Korotkov K, Nevo Y, Eyras E, Sperling R, Stamm S. 2016. Dual function of C/D box small nucleolar RNAs in rRNA modification and alternative pre-mRNA splicing. *Proc Natl Acad Sci U S A* **113**: E1625–E1634.

Falaleeva M, Welden JR, Duncan MJ, Stamm S. 2017. C/D-box snoRNAs form methylating and non-methylating ribonucleoprotein complexes: Old dogs show new tricks. *BioEssays News Rev Mol Cell Dev Biol* **39**.

Fang X, Yang D, Luo H, Wu S, Dong W, Xiao J, Yuan S, Ni A, Zhang K-J, Liu X-Y, et al. 2017. SNORD126 promotes HCC and CRC cell growth by activating the PI3K–AKT pathway through FGFR2. *J Mol Cell Biol* **9**: 243–255.

Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR. 2003. Rfam: an RNA family database. *Nucleic Acids Res* **31**: 439–441.

Hebras J, Krogh N, Marty V, Nielsen H, Cavaillé J. 2019. Developmental changes of rRNA ribose methylations in the mouse. *RNA Biol* **17**: 150–164.

Hebras J, Marty V, Personnaz J, Mercier P, Krogh N, Nielsen H, Aguirrebengoa M, Seitz H, Pradere J-P, Guiard BP, et al. 2020. Reassessment of the involvement of Snord115 in the serotonin 2c receptor pathway in a genetically relevant mouse model. *eLife* **9**.

Huang C, Shi J, Guo Y, Huang W, Huang S, Ming S, Wu X, Zhang R, Ding J, Zhao W, et al. 2017. A snoRNA modulates mRNA 3’ end processing and regulates the expression of a subset of mRNAs. *Nucleic Acids Res* **45**: 8647–8660.

Jorjani H, Kehr S, Jedlinski DJ, Gumienny R, Hertel J, Stadler PF, Zavolan M, Gruber AR. 2016. An updated human snoRNAome. *Nucleic Acids Res* **44**: 5068–5082.

Kehr S, Bartschat S, Stadler PF, Tafer H. 2011. PLEXY: efficient target prediction for box C/D snoRNAs. *Bioinformatics* **27**: 279–280.

Kishore S, Khanna A, Zhang Z, Hui J, Balwierz PJ, Stefan M, Beach C, Nicholls RD, Zavolan M, Stamm S. 2010. The snoRNA MBII-52 (SNORD 115) is processed into smaller RNAs and regulates alternative splicing. *Hum Mol Genet* **19**: 1153–1164.

Kishore S, Stamm S. 2006. The snoRNA HBII-52 Regulates Alternative Splicing of the Serotonin Receptor 2C. *Science* **311**: 230–232.

Kiss T. 2001. Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs. *EMBO J* **20**: 3617–3622.

Kiss-László Z, Henry Y, Bachellerie J-P, Caizergues-Ferrer M, Kiss T. 1996. Site-Specific Ribose Methylation of Preribosomal RNA: A Novel Function for Small Nucleolar RNAs. *Cell* **85**: 1077–1088.

Lemus-Diaz N, Ferreira RR, Bohnsack KE, Gruber J, Bohnsack MT. 2020. The human box C/D snoRNA U3 is a miRNA source and miR-U3 regulates expression of sortin nexin 27. *Nucleic Acids Res* **48**: 8074–8089.

Lestrade L, Weber MJ. 2006. snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res* **34**: D158-162.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. 2011. ViennaRNA Package 2.0. *Algorithms Mol Biol AMB* **6**: 26.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 550.

Lu Z, Gong J, Zhang QC. 2018. PARIS: Psoralen Analysis of RNA Interactions and Structures with High Throughput and Resolution. *Methods Mol Biol Clifton NJ* **1649**: 59–84.

Lu Z, Zhang QC, Lee B, Flynn RA, Smith MA, Robinson JT, Davidovich C, Gooding AR, Goodrich KJ, Mattick JS, et al. 2016. RNA duplex map in living cells reveals higher order transcriptome structure. *Cell* **165**: 1267–1279.

Mann M, Wright PR, Backofen R. 2017. IntaRNA 2.0: enhanced and customizable prediction of RNA–RNA interactions. *Nucleic Acids Res* **45**: W435–W439.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. 2011. Scikit-learn: Machine Learning in Python. *J Mach Learn Res* **12**: 2825–2830.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinforma Oxf Engl* **26**: 841–842.

Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, Vilo J. 2019. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res* **47**: W191–W198.

Schaffer JE. 2020. Death by lipids: The role of small nucleolar RNAs in metabolic stress. *J Biol Chem* **295**: 8628–8635.

Schönberger B, Schaal C, Schäfer R, Voß B. 2018. RNA interactomics: recent advances and remaining challenges. *F1000Research* **7**: F1000 Faculty Rev-1824.

Scott MS, Ono M, Yamada K, Endo A, Barton GJ, Lamond AI. 2012. Human box C/D snoRNA processing conservation across multiple cell types. *Nucleic Acids Res* **40**: 3676–3688.

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

Soeno Y, Taya Y, Stasyk T, Huber LA, Aoba T, Hüttenhofer A. 2010. Identification of novel ribonucleo-protein complexes from the brain-specific snoRNA MBII-52. *RNA* **16**: 1293–1300.

Van Nostrand EL, Freese P, Pratt GA, Wang X, Wei X, Xiao R, Blue SM, Chen J-Y, Cody NAL, Dominguez D, et al. 2020. A large-scale binding and functional map of human RNA-binding proteins. *Nature* **583**: 711–719.

Van Nostrand EL, Pratt GA, Shishkin AA, Gelboin-Burkhart C, Fang MY, Sundararaman B, Blue SM, Nguyen TB, Surka C, Elkins K, et al. 2016. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat Methods* **13**: 508–514.

Vaquero-Garcia J, Barrera A, Gazzara MR, González-Vallinas J, Lahens NF, Hogenesch JB, Lynch KW, Barash Y. 2016. A new view of transcriptome complexity and regulation through the lens of local splicing variations. *eLife* **5**: e11752.

Vitali P, Basyuk E, Le Meur E, Bertrand E, Muscatelli F, Cavaillé J, Huttenhofer A. 2005. ADAR2-mediated editing of RNA substrates in the nucleolus is inhibited by C/D small nucleolar RNAs. *J Cell Biol* **169**: 745–753.

Weinstein LB, Steitz JA. 1999. Guided tours: from precursor snoRNA to functional snoRNP. *Curr Opin Cell Biol* **11**: 378–384.

Yang SL, DeFalco L, Anderson DE, Zhang Y, Aw AJ, Lim SY, Ni LX, Tan KY, Zhang T, Chawla T, et al. 2021. *Comprehensive mapping of SARS-CoV-2 interactions in vivo reveals functional virus-host interactions*. https://www.biorxiv.org/content/10.1101/2021.01.17.427000v1 (Accessed August 31, 2021).

Yu F, Bracken CP, Pillman KA, Lawrence DM, Goodall GJ, Callen DF, Neilsen PM. 2015. p53 Represses the Oncogenic Sno-MiR-28 Derived from a SnoRNA. *PloS One* **10**: e0129190.

Zhang D, Zhou J, Gao J, Wu R-Y, Huang Y-L, Jin Q-W, Chen J-S, Tang W-Z, Yan L-H. 2019. Targeting snoRNAs as an emerging method of therapeutic development for cancer. *Am J Cancer Res* **9**: 1504–1516.

Zhong F, Zhou N, Wu K, Guo Y, Tan W, Zhang H, Zhang X, Geng G, Pan T, Luo H, et al. 2015. A SnoRNA-derived piRNA interacts with human interleukin-4 pre-mRNA and induces its decay in nuclear exosomes. *Nucleic Acids Res* **43**: 10474–10491.

Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31**: 3406–3415.

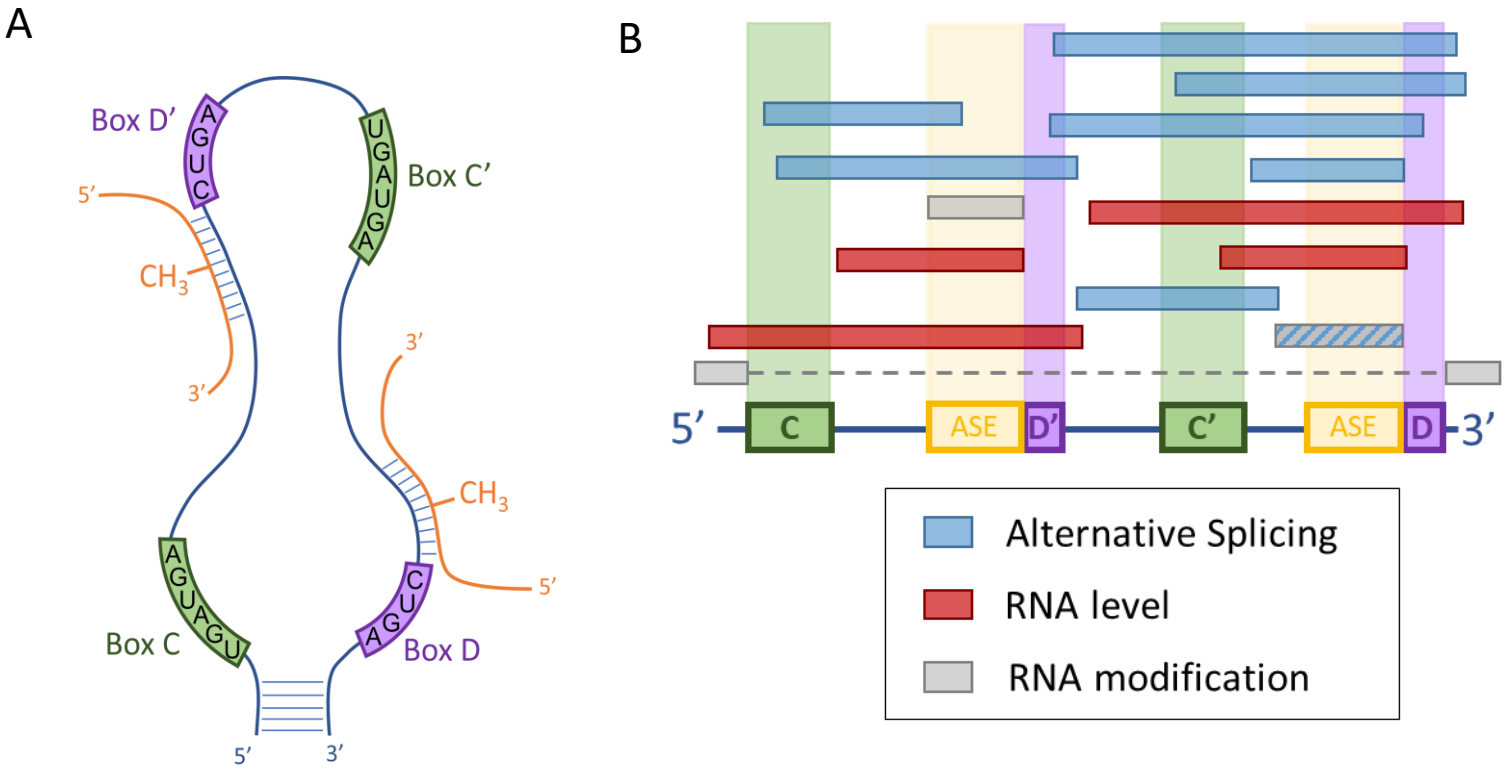


Figure 1
Deschamps-Francoeur et al. 2021

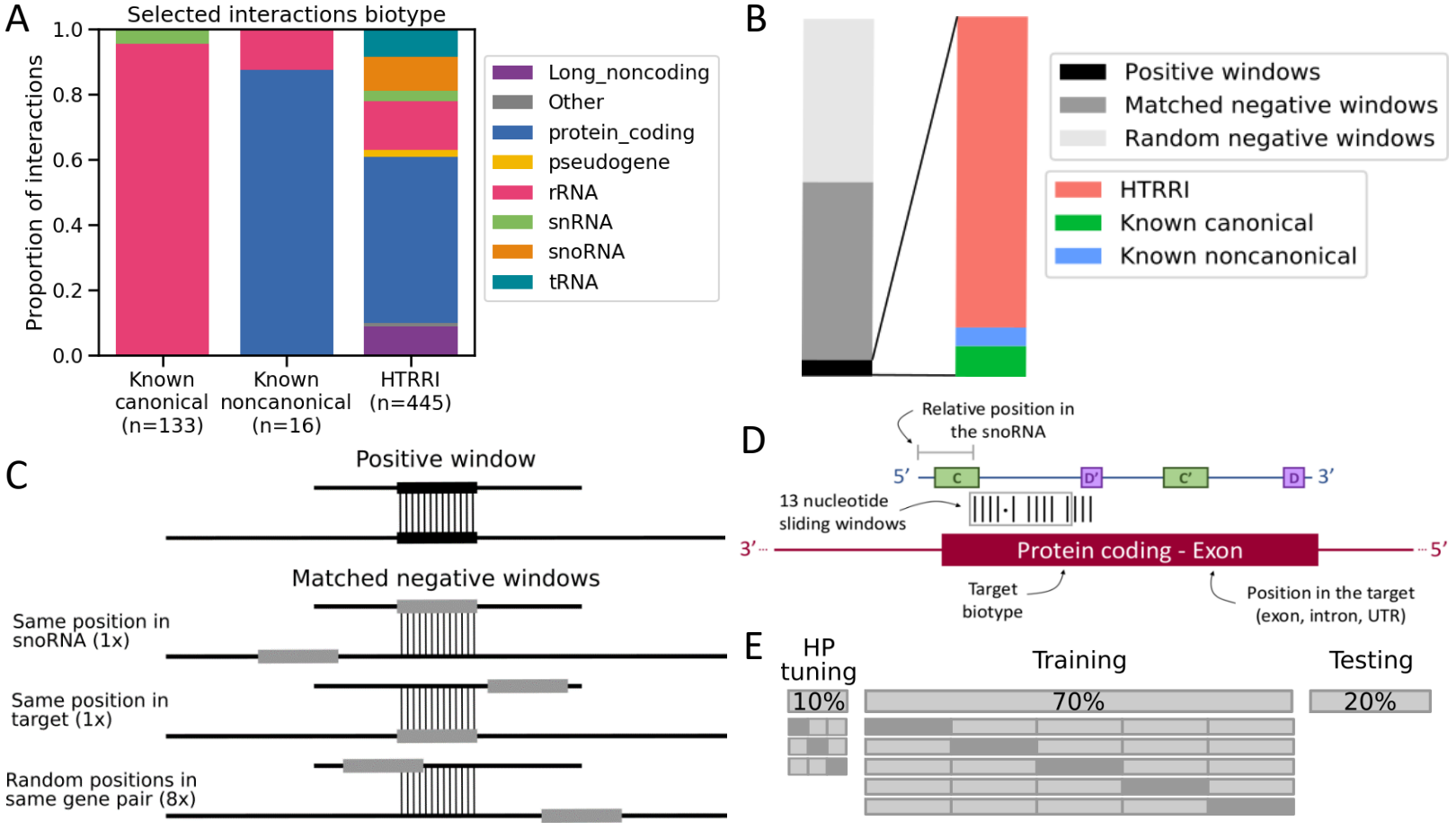
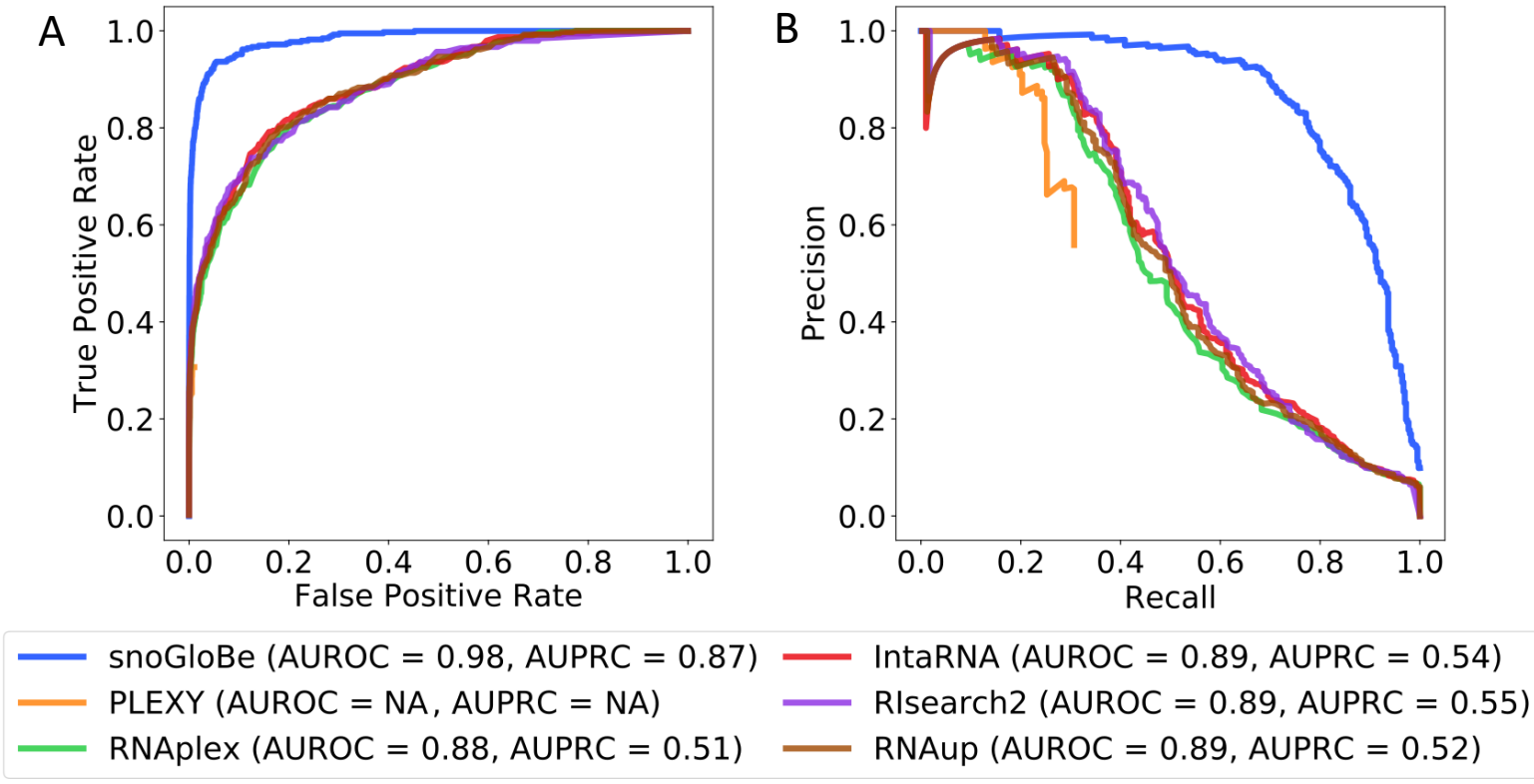


Figure 2
Deschamps-Francoeur et al. 2021



C

	snoGloBe	RNAplex	IntaRNA	RIsearch2	RNAup	PLEXY
Threshold used	0.503	-14.40	-13.51	-12.35	-13.87	-9.45
True positives (% of positives)	**275 (69.8)**	108 (27.4)	113 (28.7)	123 (31.2)	114 (28.9)	40 (19.9)
False positives (% of negatives)	30 (0.4)	12 (0.2)	12 (0.2)	14 (0.2)	13 (0.2)	**4 (0.1)**
False negatives (% of positives)	**119 (30.2)**	286 (72.6)	281 (71.3)	271 (68.8)	280 (71.1)	161 (80.1)
True negatives (% of negatives)	7910 (99.6)	7928 (99.8)	7928 (99.8)	7926 (99.8)	7927 (99.8)	**4643 (99.9)**
Matthews correlation coefficient	**0.78**	0.49	0.50	0.52	0.50	0.42
Accuracy	**0.98**	0.96	0.96	0.97	0.96	0.97
Precision	0.90	0.90	0.90	0.90	0.90	0.91
F1-score	**0.79**	0.42	0.44	0.46	0.44	0.33

D

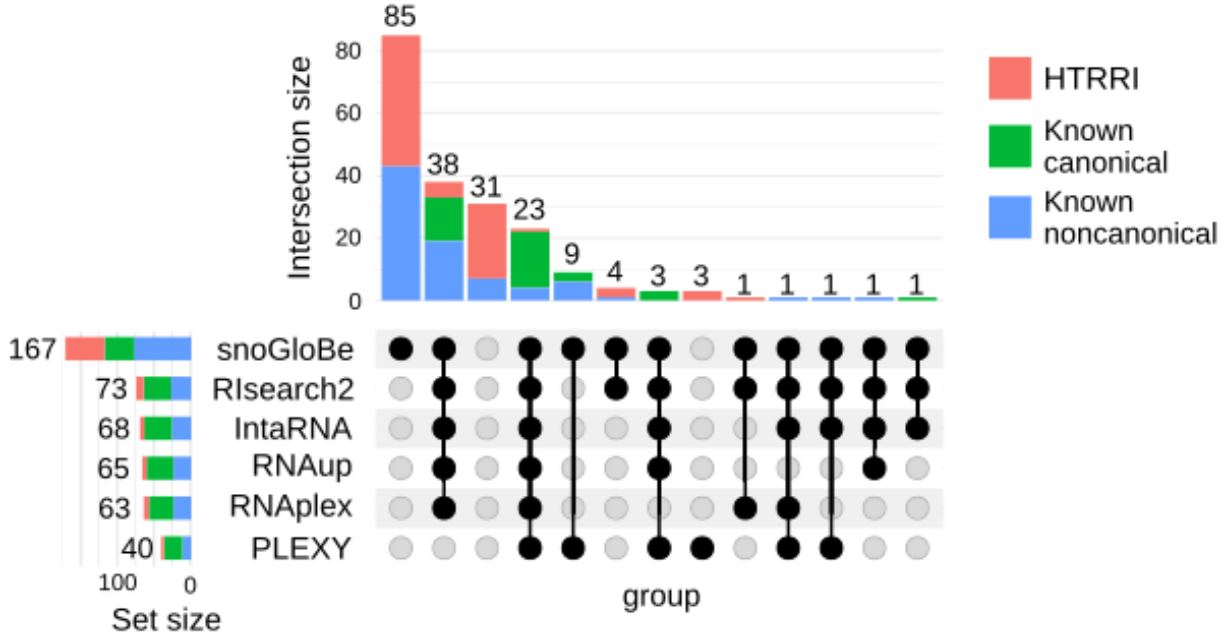


Figure 3
Deschamps-Francoeur et al. 2021

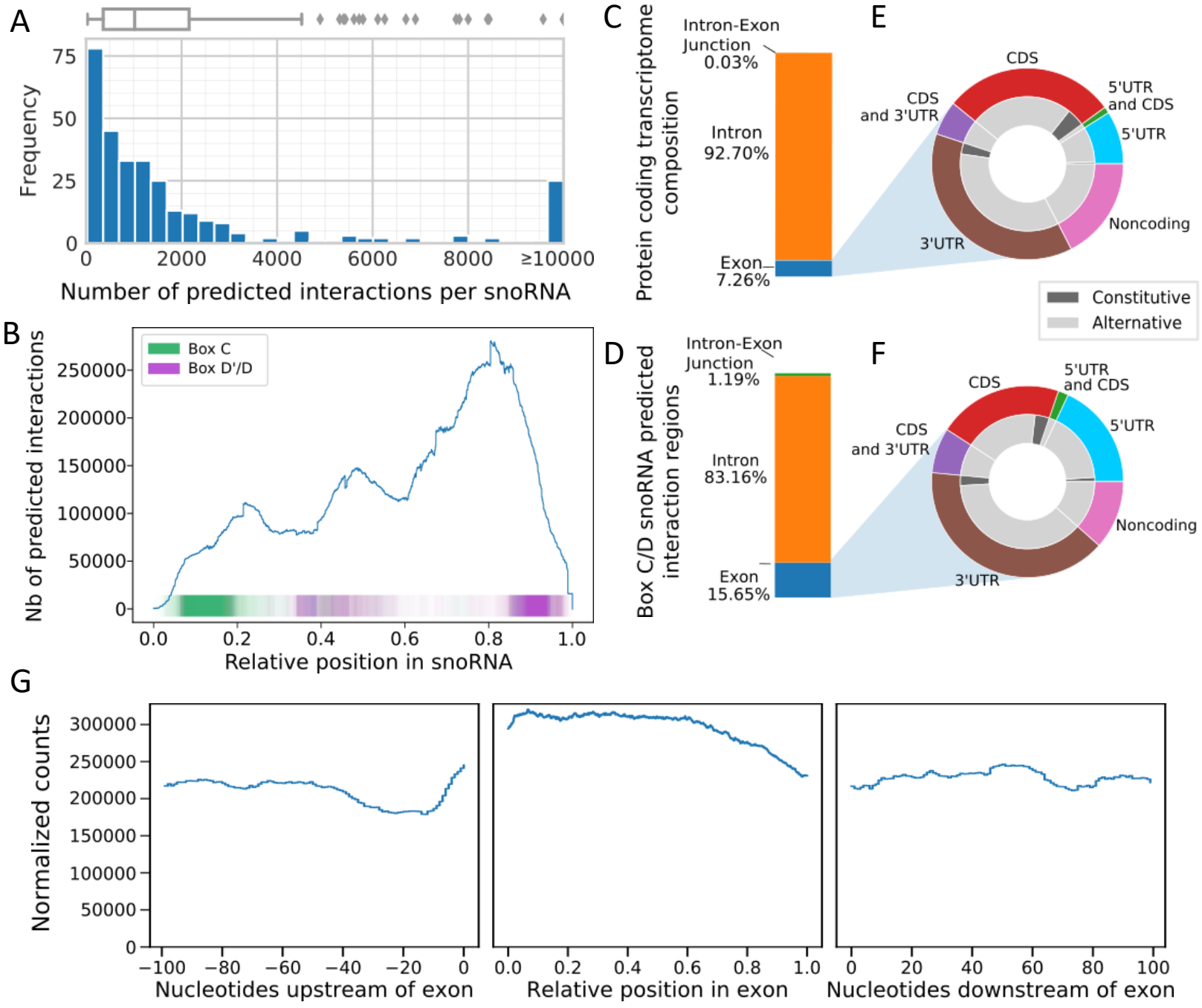


Figure 4
Deschamps-Francoeur et al. 2021

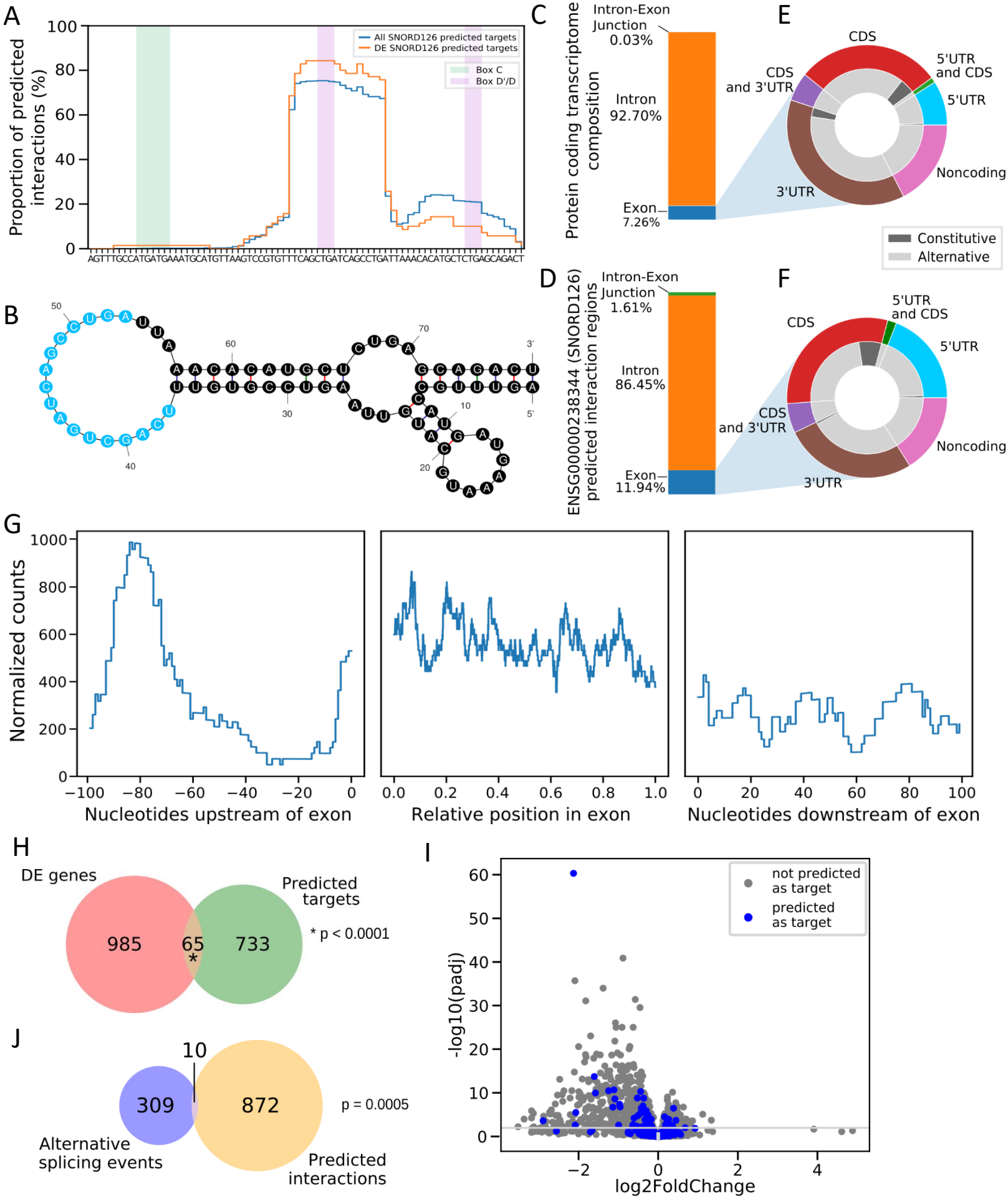
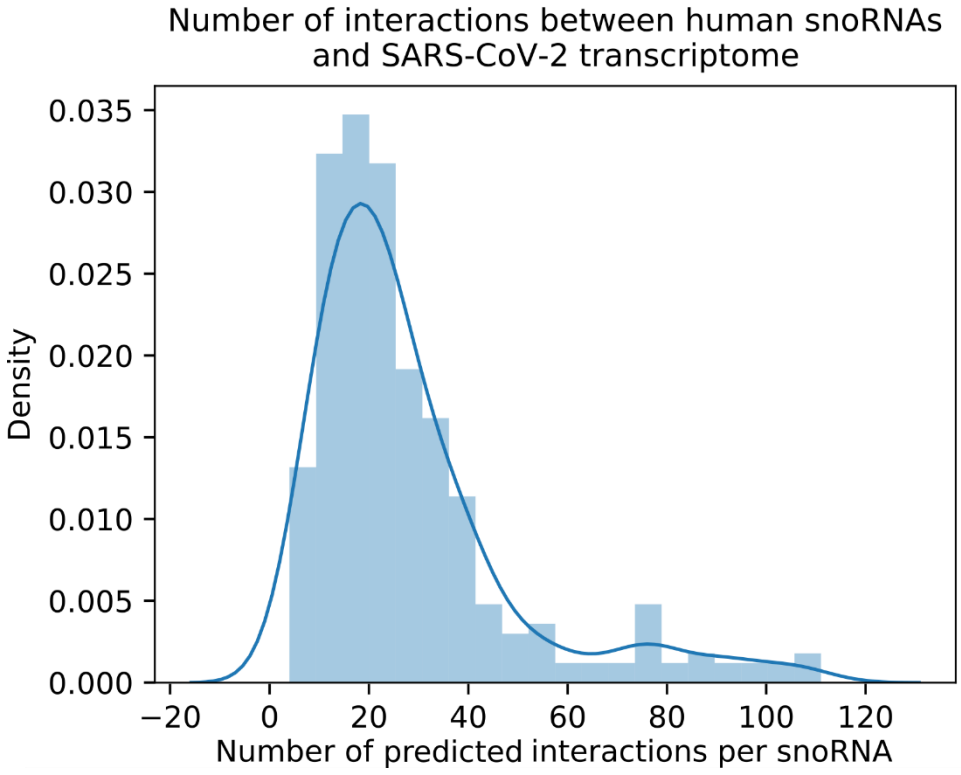


Figure 5
Deschamps-Francoeur et al. 2021

A



B

Figure 6
Deschamps-Francoeur et al. 2021

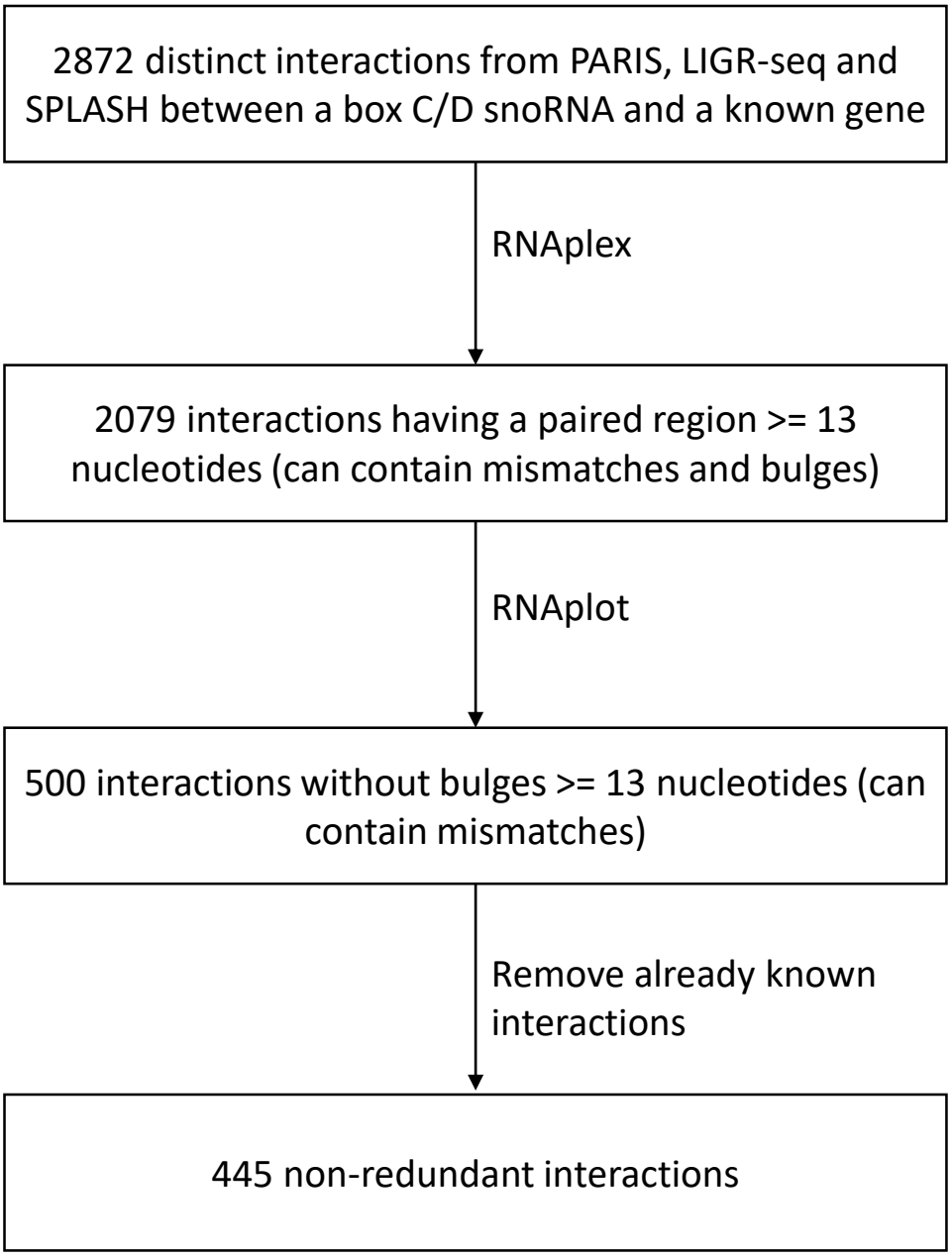


Figure S1: Overview of the filtering procedure starting with all distinct interactions detected in at least one HTRRI dataset involving a snoRNA to obtain the final HTRRI formatted datasets. The number of interactions remaining at each step is indicated.

Tool	Options used	Version	Ref
snoGloBe	-t 0.505	0.1.0	This paper
PLEXY	-e -9.29	N/A	Kehr et al., 2011
RNAplex	-e -14.51	2.4.14	Lorenz et al., 2011
RNAup	-e -13.97	2.4.14	Lorenz et al., 2011
RIsearch2	-e -12.49 -s 1	2.1	Alkan et al., 2017
IntaRNA	-e -13.30 --noSeed	3.1.1	Mann et al., 2017

Figure S2: List of tools compared for the prediction of snoRNA-RNA interactions as well as their parameter values used, version and paper reference.

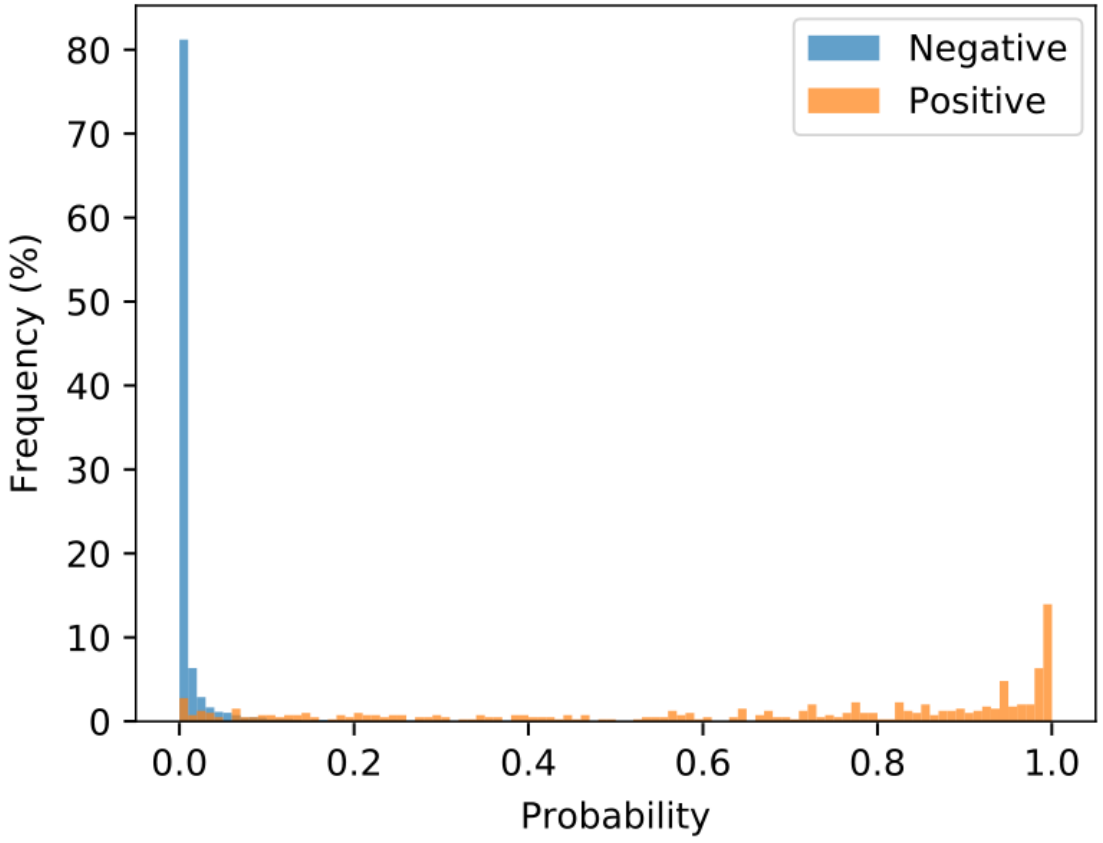


Figure S3: Distribution of the scores output by snoGloBe for the positive and negative examples in the test set.

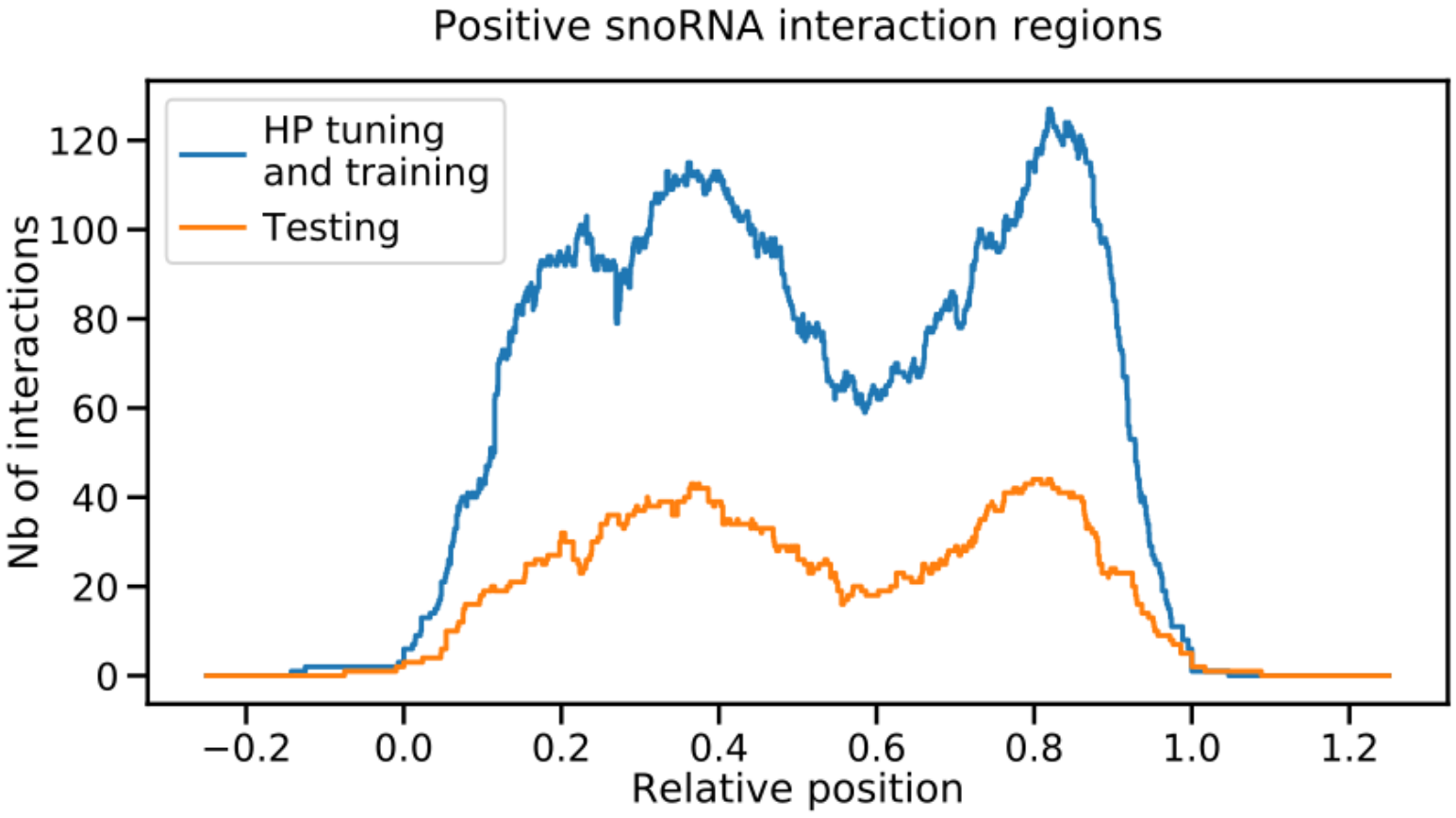


Figure S4: Profiles representing the number of predicted interactions from the hyperparameter tuning and training sets (blue) or the test set (orange) that involve specific positions in the snoRNA, for all snoRNAs considered simultaneously.

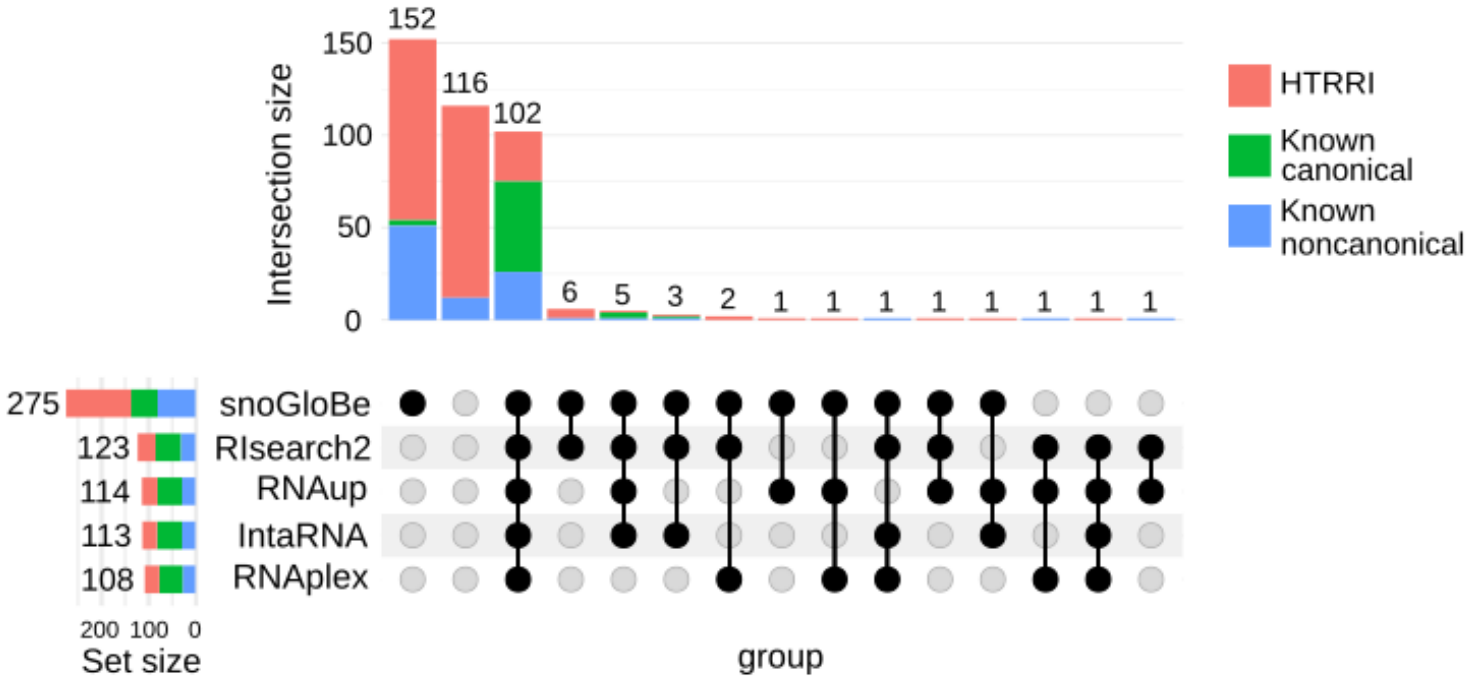


Figure S5: Upset plot indicating the number of correctly predicted interactions for all predictors compared, considering all the box C/D snoRNAs. The color legend on the right indicates the different interaction types.

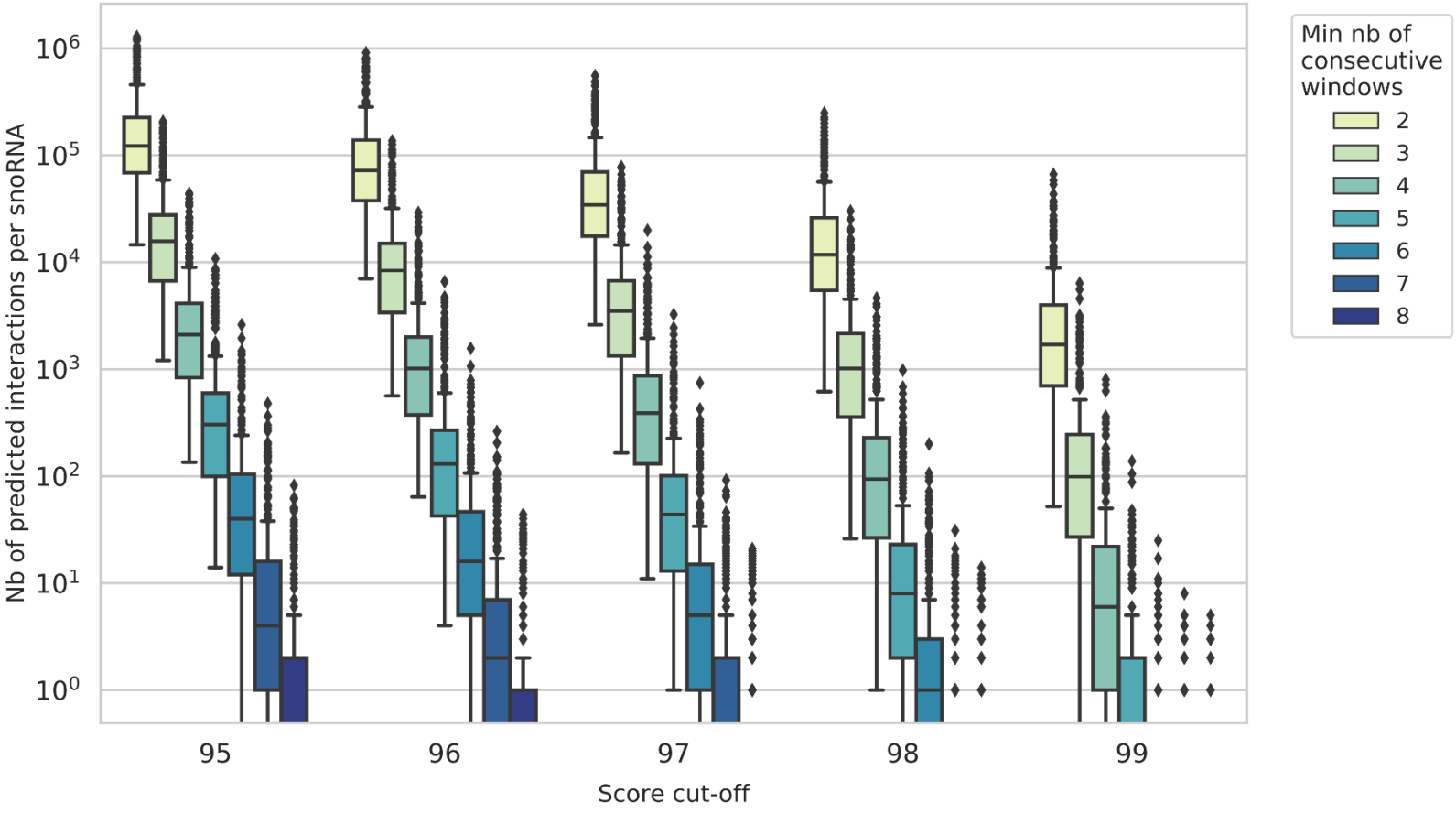


Figure S6: Boxplots depicting the distributions of the number of predicted interactions as a function of the parameter values used. The parameters considered are the number of predicted interactions per snoRNA (y axis), the score cut-off (x axis) and the minimum number of consecutive windows (different colors, with legend on the right).

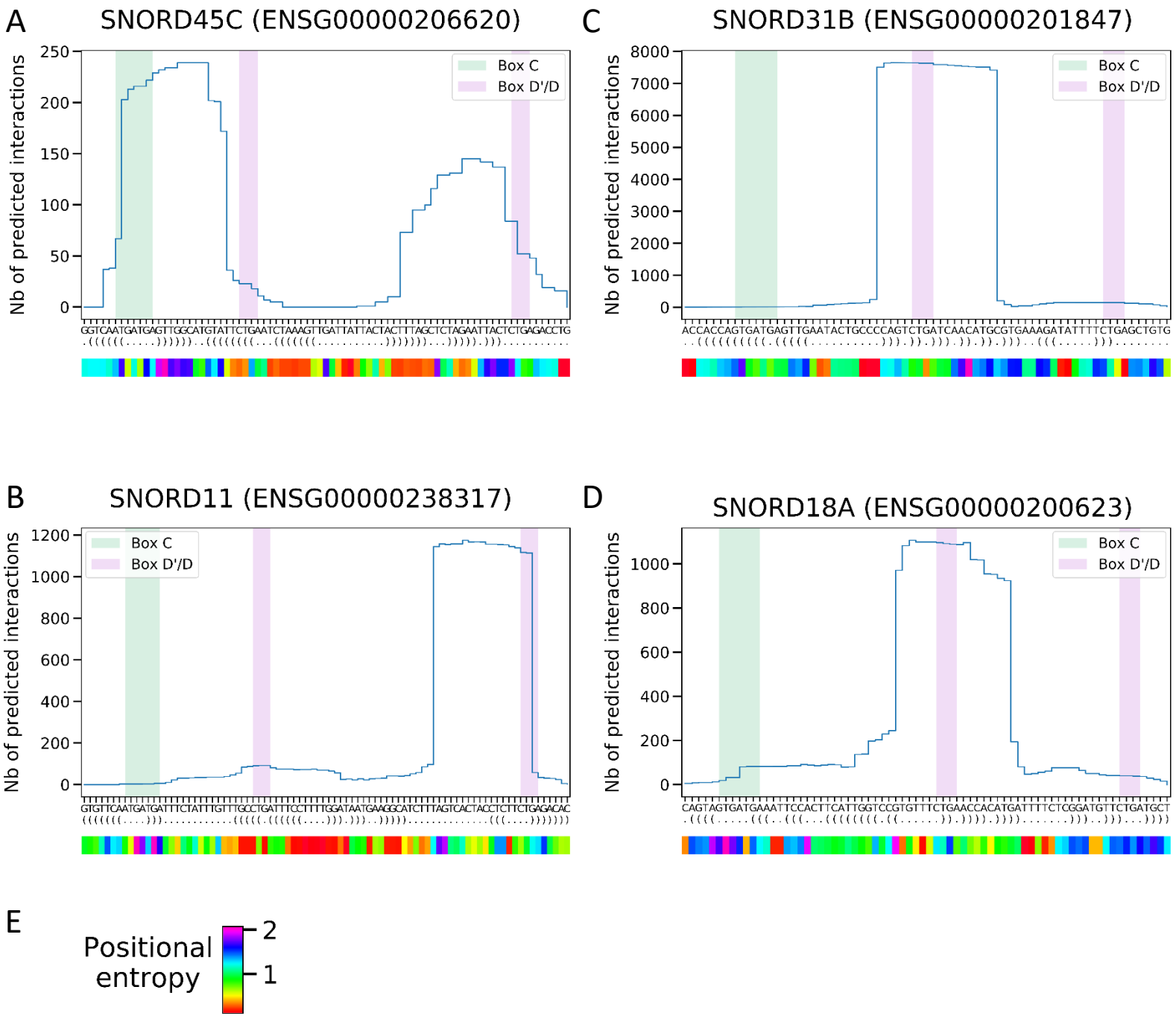


Figure S7: Profiles measuring the number of predicted interactions involving each position of the snoRNAs SNORD45C (A), SNORD11 (B), SNORD31B (C) and SNORD18A (D). The green and pink highlighting in the profiles represent respectively the positions of the C and D'/D boxes. The predicted folding of each snoRNA, as predicted by RNAplot, is shown using the dot-bracket format. The panel below each profile represents the positional entropy predicted by replot from the ViennaRNA package for each position of the snoRNA. The color legend for the position entropy score is given in E.

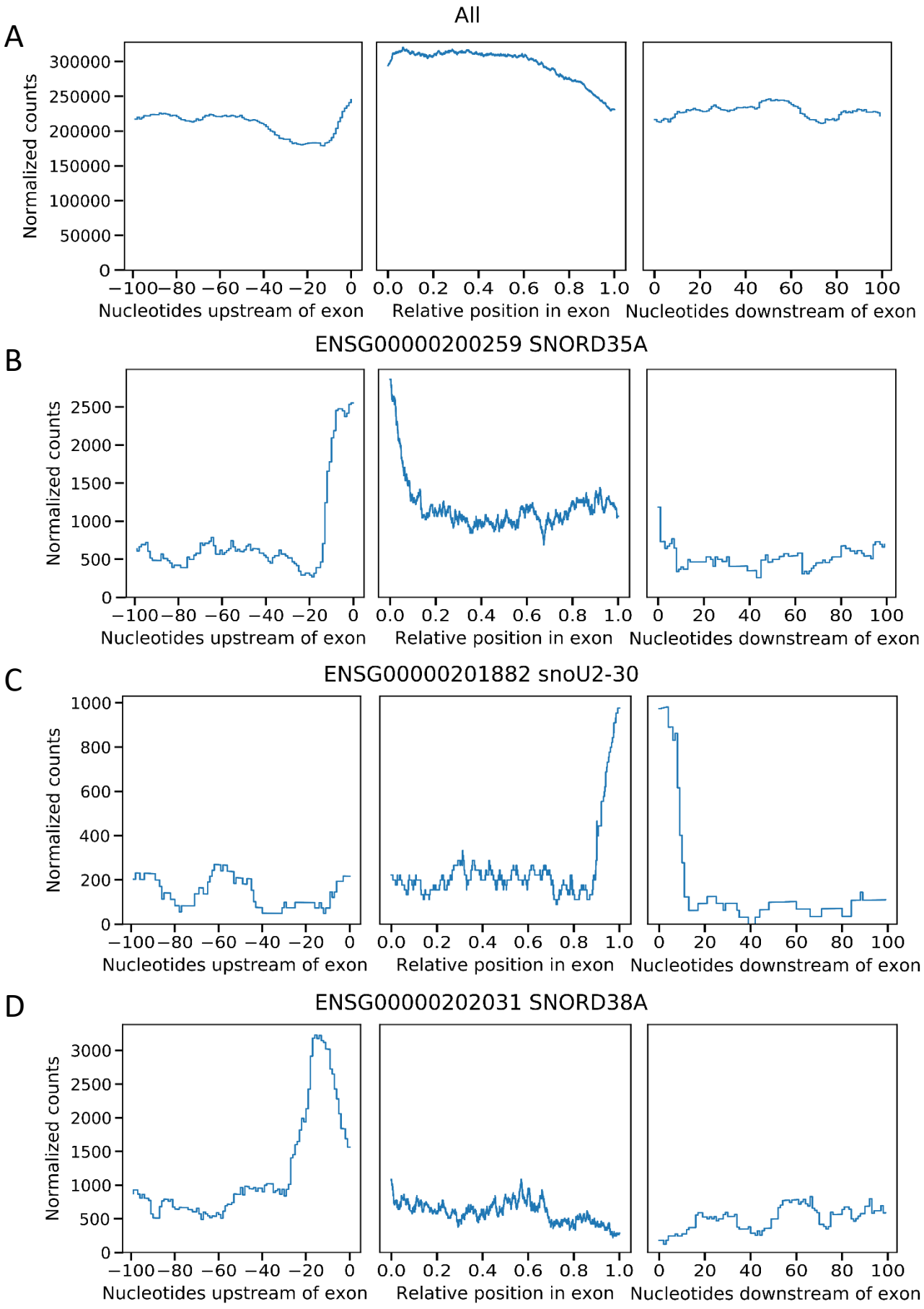
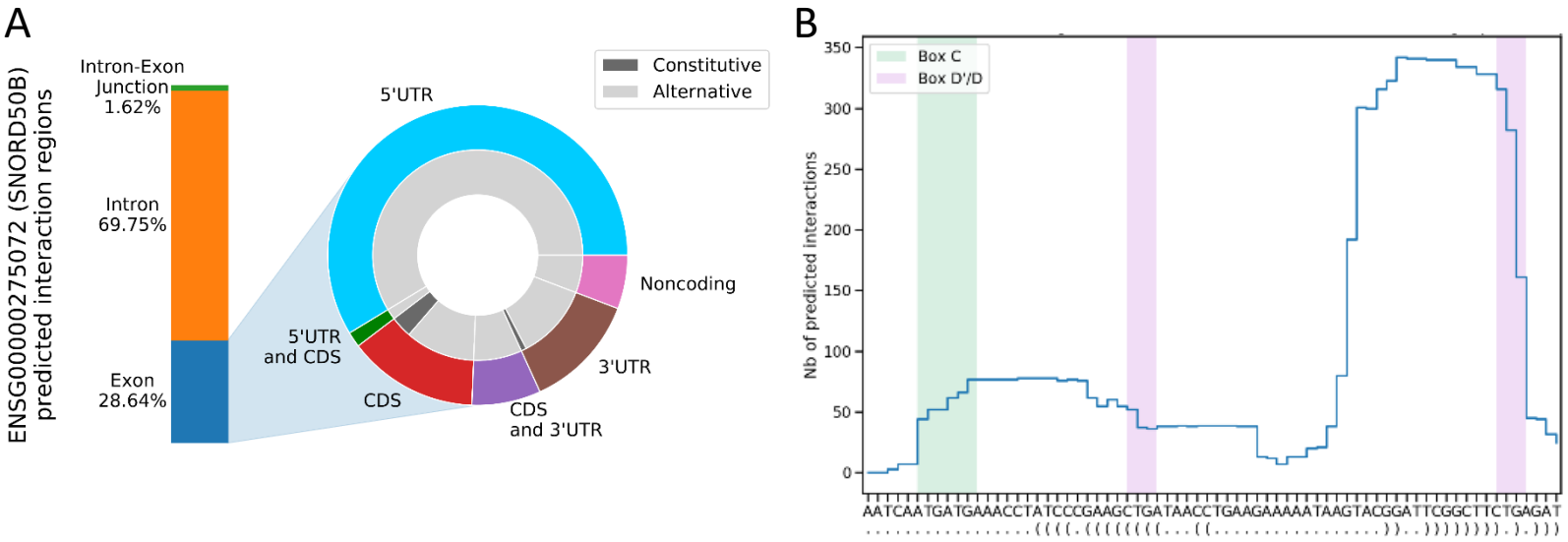
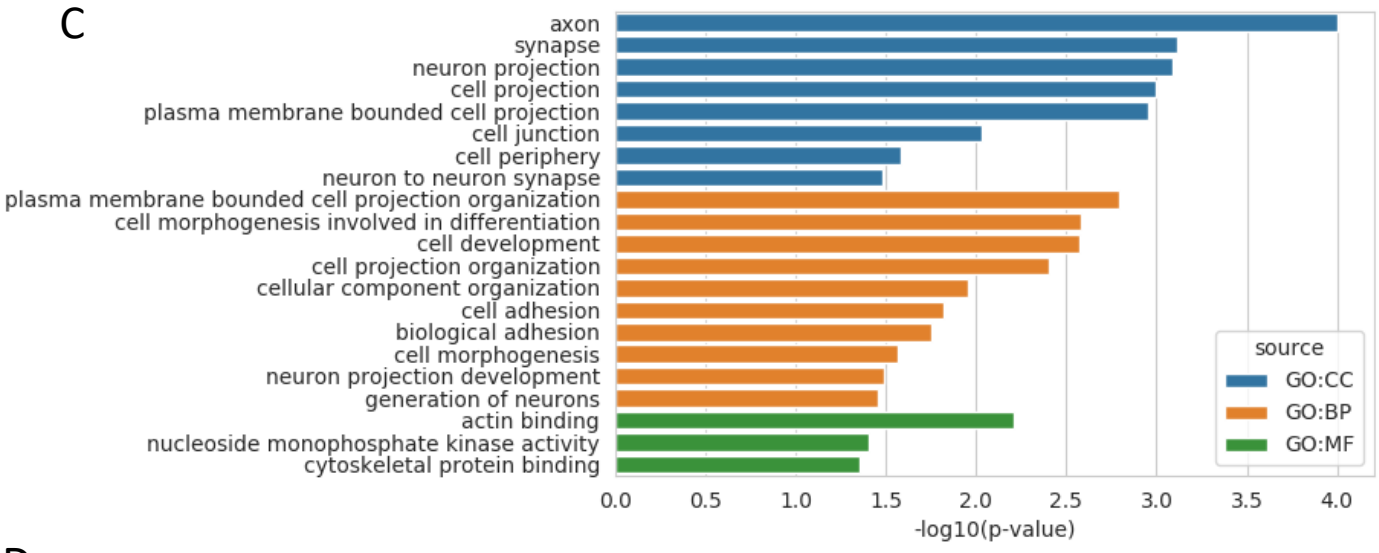


Figure S8: Profiles measuring the number of predicted interactions as a function of position with respect to the termini of introns and exons for all targets of all snoRNAs (A), SNORD35A (B), snoU2-30 (C) and SNORD38A (D).





D

RBP	Number of predicted interactions overlapping binding sites	Significant target overlap (p-value < 10^{-5})	RBP binds snoRNA	snoRNA binds RBP pre-mRNA
DDX3X	43	X		
NCBP2	29	X	X	
GRWD1	25	X		X
BUD13	21	X	X	
FTO	19	X		X

Figure S9 SNORD50B displays a positional enrichment of binding on functionally related targets which are targeted by related RBPs. (A) SNORD50B targets are enriched in exons and more specifically in 5' UTRs compared to the transcriptome distribution (Fig. 4 C and E) as shown using a stacked bar plot showing the distribution of its targets across introns and exons (left) with a doughnut plot showing the breakdown for different exonic elements (right). (B) SNORD50B binds its targets using mainly its box D ASE, as shown with a positional profile measuring the number of targets each position of the snoRNA is predicted to bind. (C) SNORD50B predicted targets are enriched in specific gene ontology terms (CC: cellular component, BP: biological process, MF: molecular function) indicated using a horizontal bargraph. (D) SNORD50B displays strong association with specific RBPs. Table showing the number of predicted SNORD50B interactions that overlap binding sites for the indicated RBPs. (All 5 RBPs indicated have a significant overlap). In addition, 2 of the RBPs bind SNORD50B according to eCLIP experiments and SNORD50B is predicted to bind the pre-mRNA of 2 of the RBPs.

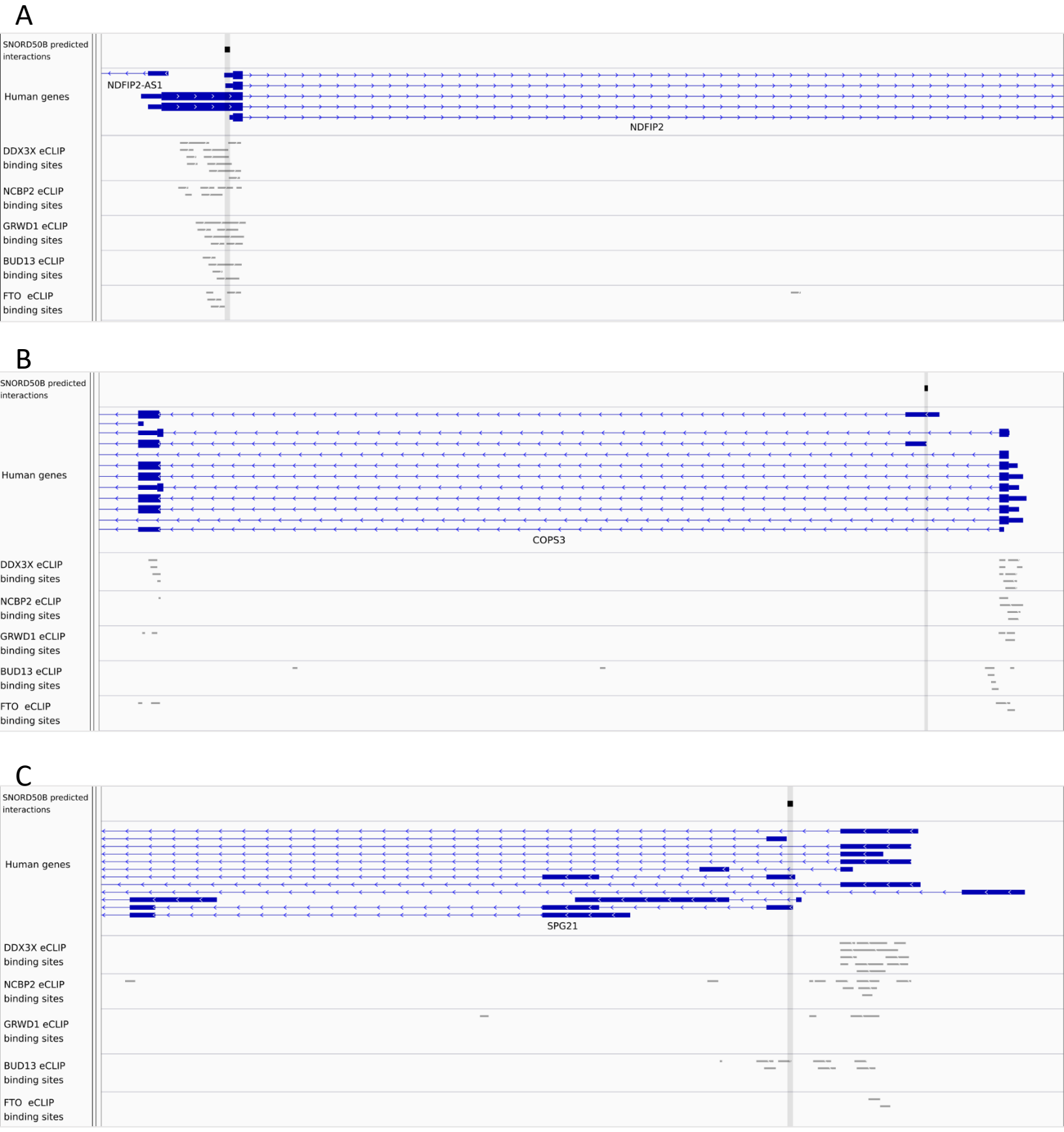
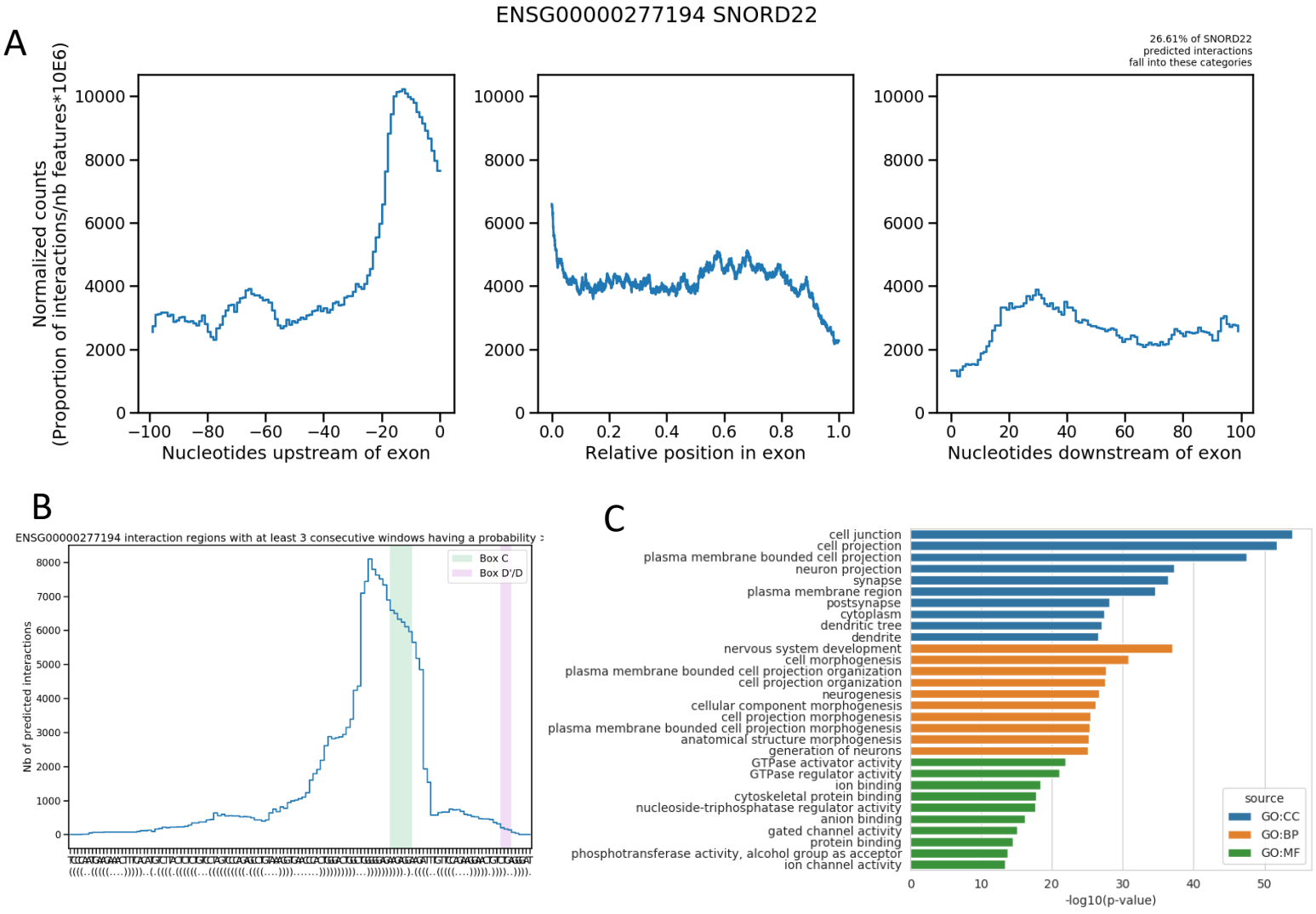


Figure S10 Examples of SNORD50B binding sites on alternative 5' UTRs. SNORD50B binding sites are enriched in 5' UTRs many of which are alternative including those in (A) NDIFP2, (B) COPS3 and (C) SPF21, as shown with genome browser screenshots. In each case, the top track displays the predicted binding position. The Human genes track indicates the architecture of the different isoforms encoded for these genes for the positional window chosen and the bottom tracks show the binding sites for the 5 RBPs indicated in Figure S9D according to ENCODE eCLIPs.



D

RBP	Number of predicted interactions overlapping binding sites	Significant target overlap (p-value < 10^{-5})	RBP binds snoRNA	snoRNA binds RBP pre-mRNA
PCBP2	363	X		
PTBP1	264	X	X	X
BUD13	256	X		
PRPF8	212	X	X	
AQR	207	X	X	

Figure S11 SNORD22 displays a positional enrichment of binding on functionally related targets which are targeted by related RBPs involved in splicing. (A) SNORD22 targets are enriched in 3' SSs of its targets as shown using a positional profile covering the last 100 nt of introns, the relative position in exons and the first 100 nt in introns. (B) SNORD22 binds its targets using a region overlapping its box C'. (C) SNORD22 predicted targets are enriched in specific gene ontology terms (CC: cellular component, BP: biological process, MF: molecular function) indicated using a horizontal bargraph. (D) SNORD22 displays strong association with specific RBPs. Table showing the number of predicted SNORD22 interactions that overlap binding sites for the indicated RBPs. (All 5 RBPs indicated have a significant overlap). In addition, 3 of the RBPs bind SNORD22 according to eCLIP experiments and SNORD22 is predicted to bind the pre-mRNA of 1 of the RBPs.

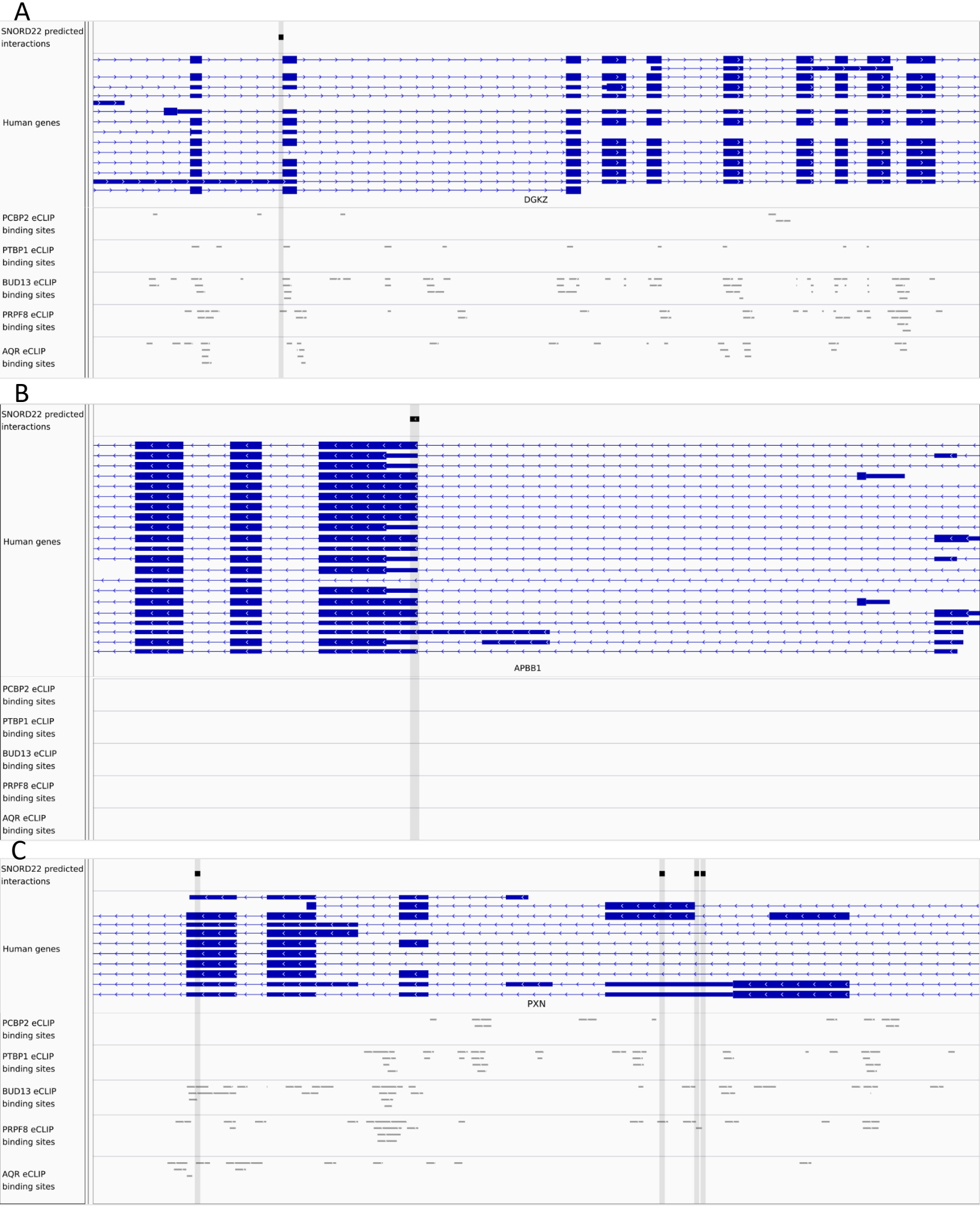
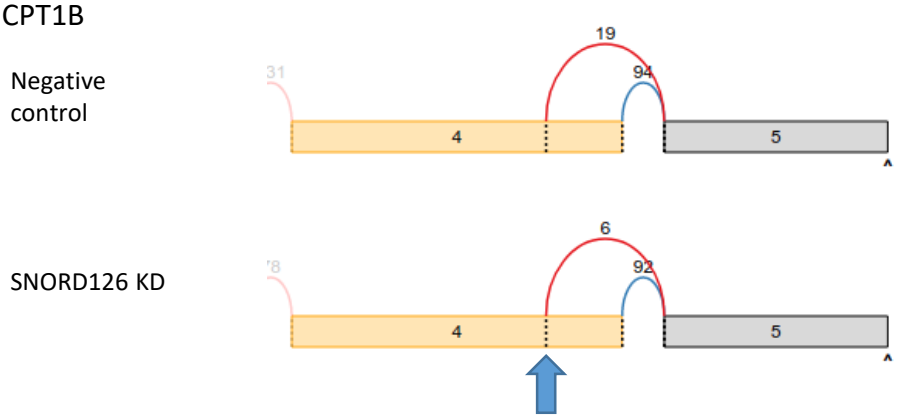


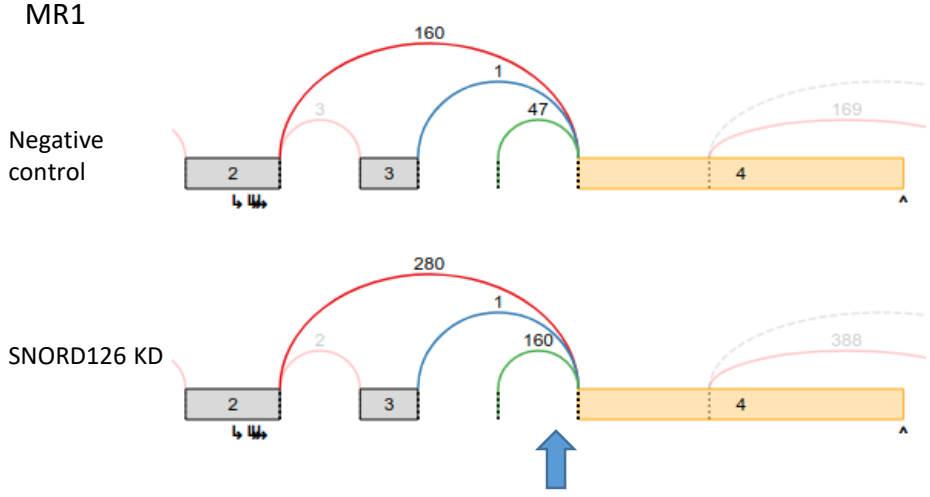
Figure S12 Examples of SNORD22 binding sites on alternative 3'SSs. SNORD22 binding sites are enriched in 3' SSs many of which are alternative including those in (A) DGKZ, (B) APBB1 and (C) PXN. Screenshots are as described in Figure S10.

A CPT1B



Delta PSI: -0.101, 0.101
Probability non-changing: 0.0004, 0.0004
Probability changing: 0.52, 0.52

B MR1



Delta PSI: -0.12, 0, 0.12
Probability non-changing: 0.0002, 0.48, 0.0001
Probability changing: 0.65, 5e-21, 0.67

Figure S13: Examples of splicing events affected by the knockdown of SNORD126. Both CPT1B (A) and MR1 (B) display differential splicing following the knockdown (KD) of SNORD126 as shown using sashimi plots. The blue arrows represent the predicted interaction region. The colored arcs represent different splice junctions with the number of reads supporting them. Statistics of the splicing event are given on the right.

ASO	Sequence
SNORD126_ASO1	5’-mC*mU*mG*mA*mA*A*C*A*C*G*G*A*C*T*T*mA*mA*mC*mA*mU*-3’
SNORD126_ASO2	5’-mA*mG*mC*mA*mU*G*T*G*T*T*T*A*A*T*C*mA*mG*mG*mC*mU*-3’
NC5	5’-mG*mC*mG*mA*mC*T*A*T*A*C*G*C*G*C*A*mA*mU*mA*mU*mG-3’

Figure S14: List of ASO sequences used for SNORD126 knockdown and negative control (NC5). * means phosphorothioate backbone, m means 2’-O-methoxyethyl.