APOBEC1 mediated C-to-U RNA editing: target sequence and trans-acting factor contribution to 177 RNA editing events in 119 murine transcripts in-vivo.

Saeed Soleymanjahi[1], Valerie Blanc[1] and Nicholas O. Davidson[1,2]

[1]Division of Gastroenterology, Department of Medicine, Washington University School of Medicine, St. Louis, MO 63105

[2]To whom communication should be addressed:

Email: nod@wustl.edu

Running title: APOBEC1 mediated C to U RNA editing

Keywords: RNA folding; A1CF; RBM47;

January 8, 2021

ABSTRACT (184 words)

Mammalian C-to-U RNA editing was described more than 30 years ago as a single nucleotide modification in APOB RNA in small intestine, later shown to be mediated by the RNA-specific cytidine deaminase APOBEC1. Reports of other examples of C-to-U RNA editing, coupled with the advent of genome-wide transcriptome sequencing, identified an expanded range of APOBEC1 targets. Here we analyze the cis-acting regulatory components of verified murine C-to-U RNA editing targets, including nearest neighbor as well as flanking sequence requirements and folding predictions. We summarize findings demonstrating the relative importance of trans-acting factors (A1CF, RBM47) acting in concert with APOBEC1. Using this information, we developed a multivariable linear regression model to predict APOBEC1 dependent C-to-U RNA editing efficiency, incorporating factors independently associated with editing frequencies based on 103 Sanger-confirmed editing sites, which accounted for 84% of the observed variance. Co-factor dominance was associated with editing frequency, with RNAs targeted by both RBM47 and A1CF observed to be edited at a lower frequency than RBM47 dominant targets. The model also predicted a composite score for available human C-to-U RNA targets, which again correlated with editing frequency.

INTRODUCTION

Mammalian C-to-U RNA editing was identified as the molecular basis for human intestinal APOB48 production more than three decades ago (Chen et al. 1987; Hospattankar et al. 1987; Powell et al. 1987). A site-specific enzymatic deamination of C6666 to U of *Apob* mRNA was originally considered the sole example of mammalian C-to-U RNA editing, occurring at a single nucleotide in a 14 kilobase transcript and mediated by an RNA specific cytidine deaminase (APOBEC1) (Teng et al. 1993). With the advent of massively parallel RNA sequencing technology we now appreciate that APOBEC1 mediated RNA editing targets hundreds of sites (Rosenberg et al. 2011; Blanc et al. 2014) mostly within 3' untranslated regions of mRNA transcripts. This expanded range of targets of C-to-U RNA editing prompted us to reexamine key functional attributes in the regulatory motifs (both cis-acting elements and trans-acting factors) that impact editing frequency, focusing primarily on data emerging from studies of mouse cell and tissue-specific C-to-U RNA editing.

Earlier studies identified RNA motifs (Davies et al. 1989) contained within a 26-nucleotide segment flanking the edited cytidine base *in vivo* (in cell lines) or within 55 nucleotides using S100 extracts from rat hepatoma cells (Bostrom et al. 1989; Driscoll et al. 1989). Those, and other studies, established that *Apob* RNA editing reflects both the tissue/cell of origin as well as RNA elements remote and adjacent to the edited base (Bostrom et al. 1989; Davies et al. 1989). A granular examination of the regions flanking the edited base in *Apob* RNA demonstrated a critical 3' sequence 6671-6681, downstream of C6666, in which mutations reduced or abolished editing activity (Shah et al. 1991). This 3' site, termed a "mooring sequence" was associated with a 27s- "editosome" complex (Smith et al. 1991), which was both necessary and sufficient for site-specific *Apob* RNA editing and editosome assembly (Backus and Smith 1991). Other *cis*-acting elements include a 5 nucleotide spacer region between the edited cytidine and the mooring sequence, and also sequences 5' of the editing site that regulate editing efficiency

(Backus and Smith 1992; Backus et al. 1994) along with AU-rich regions both 5' and 3' of the edited cytidine that together function in concert with the mooring sequence (Hersberger and Innerarity 1998).

Advances in our understanding of physiological *Apob* RNA editing emerged in parallel from both the delineation of key RNA regions (summarized above) and also with the identification of components of the *Apob* RNA editosome (Sowden et al. 1996). APOBEC1, the catalytic deaminase (Teng et al. 1993) is necessary for physiological C-to-U RNA editing *in vivo* (Hirano et al. 1996) and *in vitro* (Giannoni et al. 1994). Using the mooring sequence of *Apob* RNA as bait, two groups identified APOBEC1 complementation factor (A1CF), an RNA-binding protein sufficient *in vitro* to support efficient editing in presence of APOBEC1 and *Apob* mRNA (Lellek et al. 2000; Mehta et al. 2000). Those findings reinforced the importance of both the mooring sequence and an RNA binding component of the editosome in promoting *Apob* RNA editing. However, while A1CF and APOBEC1 are sufficient to support *in vitro Apob* RNA editing, neither heterozygous (Blanc et al. 2005) or homozygous genetic deletion of *A1cf* impaired *Apob* RNA editing *in vivo* in mouse tissues (Snyder et al. 2017), suggesting that an alternate complementation factor was likely involved. Other work identified a homologous RNA binding protein, RBM47, that functioned to promote *Apob* RNA editing both *in vivo* and *in vitro* (Fossat et al. 2014), and more recent studies utilizing conditional, tissue-specific deletion of *A1cf* and *Rbm47* indicate that both factors play distinctive roles in APOBEC1-mediated C-to-U RNA editing, including *Apob* as well as a range of other APOBEC1 targets (Blanc et al. 2019).

These findings together establish important regulatory roles for both *cis*-acting elements and *trans*-acting factors in C-to-U mRNA editing. However, the majority of studies delineating *cis*-acting elements reflect earlier, *in vitro* experiments using *ApoB* mRNA and relatively little is known regarding the role of *cis*-acting elements in tissue-specific C-to-U RNA editing of other transcripts, *in vivo*. Here we use statistical modeling to investigate the independent roles of

candidate regulatory factors in mouse C-to-U mRNA editing using data from *in vivo* studies from over 170 editing sites in 119 transcripts (Meier et al. 2005; Rosenberg et al. 2011; Gu et al. 2012; Blanc et al. 2014; Rayon-Estrada et al. 2017; Snyder et al. 2017; Blanc et al. 2019; Kanata et al. 2019). We also examined these regulatory factors in known human mRNA targets (Chen et al. 1987; Powell et al. 1987; Skuse et al. 1996; Mukhopadhyay et al. 2002; Grohmann et al. 2010; Schaefermeier and Heinze 2017).

RESULTS

Descriptive data

177 C-to-U RNA editing sites were identified based on eight studies that met inclusion and exclusion criteria (Meier et al. 2005; Rosenberg et al. 2011; Gu et al. 2012; Blanc et al. 2014; Rayon-Estrada et al. 2017; Snyder et al. 2017; Blanc et al. 2019; Kanata et al. 2019), representing 119 distinct RNA editing targets. 84% (100/119) of RNA targets were edited at one chromosomal location (Figure 1C) and 75% (89/119) of mRNA targets were edited at both a single chromosomal location and also within a single tissue (Figure 1D). The majority of editing sites occur in the 3` untranslated region (142/177; 80%), with exonic editing sites the next most abundant subgroup (28/177; 16%, Figure 1E). Chromosome X harbors the highest number of editing sites (18/177; 10%), followed by chromosomes 2 and 3 (15/177; 8.5% for both, Supplemental Figure 1). 103/177 editing sites were confirmed by Sanger sequencing, with a mean editing frequency of 37 ± 22%.

Base content of sequences flanking edited and mutated cytidines

AU content was enriched (~87%) in nucleotides both immediately upstream and downstream of the edited cytidine across mouse RNA editing targets (Figure 2A and 2C). The average AU content across the region 10 nucleotides upstream to 20 nucleotides downstream of the edited cytidine was ~70% (60 - 87%). Because APOBEC1 has been shown to be a DNA mutator (Harris et al. 2003; Wolfe et al. 2019; Wolfe et al. 2020), we determined the AU content of the mutated deoxycytidine region flanking human DNA targets (Nik-Zainal et al. 2012) to be ~66% at a site one nucleotide downstream of the edited base (Figure 2B, C). The average AU content in the sequence 10 nucleotides upstream and 10 nucleotides downstream of mutated deoxycytidines is 59% (57-66.0%). The average AU content was 90% and 80% in nucleotides immediately upstream and downstream, respectively, of the targeted deoxycytidine in a

subgroup of over 700 DNA editing events of the C to T type (Nik-Zainal et al. 2012), which is closer to the distribution found in C to U RNA editing targets. These features suggest that AU enrichment is an important component to editing function of APOBEC1 on both RNA and DNA targets, especially for the C/dC to U/dT change.

Factors influencing editing frequency

Regulatory-spacer-mooring cassette: We observed no significant associations between editing frequency and mismatches in motif A (r=-0.05, P=.46) or motif B (r=-0.1, P=.20) (Supplemental Figure 2), while mismatches in motif C and D negatively impacted editing frequency (r=-0.24, P=.001) (motif D r=-0.20, P=.008, Figure 3B). AU content of motif B showed a trend towards negative association with editing frequency (r=-0.13, P=.08 Figure 3C), but AU contents of motifs A (r=0.06, P=.4), C (r=-0.02, P=.8), and D (r=-0.02, P=.78) did not impact editing frequency (Supplemental Figure 2). The abundance of G in motif C (r=0.17, P=.02), abundance of C in motif B (r=0.13, P=.08), and G/C fraction in motif C (r=0.14, P=.04) showed either significance or a trend to associations with editing frequency. The spacer sequence averaged 5 ± 4 nucleotides, ranging from 0 to 20, with trend of association between length and editing frequency (r=-0.14, P=.09). The mean spacer sequence AU content was 73 ± 23%, with no association between editing frequency and AU content (r=-0.1, P=.2, Supplemental Figure 3). However, G abundance (r=-0.23, P=.01) and G/C fraction (r=-0.20, P=.03) of spacer showed significant associations with editing frequency in Sanger-confirmed targets. The mean number of mismatches in the first 4 nucleotides of the spacer sequence was 2.5 ± 1 with higher number of mismatches exerting a significant negative impact on editing frequency (r=-0.24, P=.01) (Figure 3D). The mean number of mismatches in the mooring sequence was 2.1 ± 1.8, ranging from 0 to 8 nucleotides. The number of mismatches showed a significant negative association with editing frequency (r=-0.30, P=.0003, Figure 3E). The base content of individual nucleotides surrounding the edited cytidine showed significant associations with editing frequency, which

was more emphasized in nucleotides closer to the edited cytidine (Figure 3F, Supplemental Table 1). Furthermore, overall AU content of downstream sequence +16 to +20 had positive impact on editing frequency (*r*=0.17, *P*=.02) (Supplemental Figure 3). However, G abundance in downstream 20 nucleotides (*r*=-0.24, *P*=.001) and G/C fraction in downstream 10 nucleotides (*r*=-0.16, *P*=.09) showed significant or a trend of significant negative associations with editing frequency in Sanger-confirmed targets.

Secondary structure: We generated a predicted secondary structure for 172 editing sites, with four subgroups based on overall structure and location of the edited cytidine: loop (C_{loop}), stem (C_{stem}), tail (C_{tail}), and non-canonical structure (NC). The majority of editing sites were in the C_{loop} subgroup (59%), followed by C_{stem} (20%), C_{tail} (13%), and NC (8%) subgroups (Figure 4A). Editing sites in the C_{tail} subgroup exhibited lower editing frequencies compared to editing sites in C_{loop} (29 ± 12 vs 41 ± 23%, *P*=.02) or C_{stem} (37 ± 21%, *P*=.04) subgroups. No significant differences were detected in other comparisons (Figure 4B). The edited cytidine was located in loop, stem, and tail of the secondary structure in 110 (64%), 38 (22%), and 24 (14%) of the edited RNAs, respectively. Editing sites with the edited cytidine within the loop exhibited significantly higher editing frequency compared to those with the edited cytidine in the tail (40 ± 24% vs 28 ± 12 %, *P*=.04). Other subgroups exhibited comparable editing frequencies (Supplemental Figure 4). The majority (78%) of editing sites contained a mooring sequence located in main stem-loop structure (Figure 4C), with the remainder located in the tail or secondary loop. Average editing efficiency was significantly higher in targets where the mooring sequence was located in the main stem-loop (Figure 4D). We also calculated the proportion of total nucleotides that constitute the main stem-loop in the secondary structure. The average ratio was 0.62 ± 0.18 ranging from 0.28 to 1 (Supplemental Table 2) with higher ratios associated with higher editing frequency of the corresponding editing site (*r*=0.20, *P*=.007) (Figure 4E). Finally, we considered the orientation of free tails in the secondary structure in

terms of length and symmetry. Symmetric free tails were observed in 59% of editing sites (Supplemental Figure 4). The length of 5' free tail showed negative association with editing frequency (r=-0.14, P=.04, Figure 4F) while no significant associations were detected between either the length of 3' tail or symmetry of tails and editing frequency (Supplemental Figure 4).

Trans-acting factors and tissue specificity: Data for relative dominance of cofactors in APOBEC1- dependent RNA editing were available for 72 editing sites for targets in small intestine or liver (Blanc et al. 2019). RBM47 was identified as the dominant factor in 60/72 (83%) sites; A1CF was the dominant factor in 5/72 (7%) editing sites with the remaining sites (7/72; 10%), exhibiting equal codominancy (Figure 5A). The average editing frequencies at editing sites revealed differences across the groups with 41 ± 20% in RBM47-dominant targets, 23 ± 14% in A1CF-dominant, and 27 ± 11% in the co-dominant group (P=.03) (Figure 5B). The majority of RNA editing targets were edited in one tissue (103/119; 86% Figure 5C), while the maximum number of tissues in which an editing target is edited (at the same site) is 5 (*Cd36*). The small intestine harbors the highest number of verified editing sites (95/177; 54%), followed by liver (31/177; 17%), and adipose tissue (19/177; 11% Figure 5D). Sites edited in brain tissue showed the highest average editing frequency (54 ± 35 %, n=11), followed by bone marrow myeloid cells (50 ± 22 %, n=4), and kidney (47 ± 29%, n=10 Figure 5E).

We then developed a multivariable linear regression model to predict APOBEC1 dependent C-to-U RNA editing efficiency, incorporating factors independently associated with editing frequencies (Table 1). This model, based on 103 Sanger-confirmed editing sites with available data for all of the parameters mentioned, accounted for 84% of variance in editing frequency of editing sites included (R^2=0.84, P<.001 Table 1). The final multivariable model revealed several factors independently associated with editing frequency, specifically the number of mismatches in mooring sequence; regulatory sequence motif D; AU content of regulatory sequence motif B; overall secondary structure for group C_{tail} vs group C_{loop}; location of mooring sequence in

secondary structure; “base content score” parameter that represents base content of the sequences flanking edited cytidine (Table 1). Removing “base content score” from the model reduced the power from R^2=0.84 to R^2=0.59. Next, we added a co-factor dominance variable and fit the model using the 72 editing sites with available data for cofactor dominance. Along with other factors mentioned above, co-factor dominance showed significant association with editing frequency (Table 1) with RNAs targeted by both RBM47 and A1CF observed to be edited at a lower frequency than RBM47 dominant targets.

Factors associated with co-factor dominance (Figure 6, Supplemental Table 3, Supplemental Figure 5), included tissue-specificity, with higher frequency of RBM47-dominant sites in small intestine compared to liver (91 vs 63%, *P*=.008) and A1CF-dominant and co-dominant editing sites more prevalent in liver. The number of mooring sequence mismatches also varied among three subgroups: 1.1 ± 1.3 in RBM47-dominant subgroup; 2.0 ± 2.5 in A1CF-dominant subgroup; and 2.9 ± 0.4 in co-dominant subgroup (*P*=.004). This was also the case regarding mismatches in the spacer: 2.4 ± 1.2 in RBM47-dominant subgroup; 2.7 ± 1.5 in A1CF-dominat subgroup; 3.8 ± 0.4 in co-dominant subgroup (*P*=.02). AU content (%) of downstream sequence +6 to +10 was higher in RBM47-dominant subgroup (P=.01). Finally, the location of the edited cytidine in secondary structure of mRNA strand was different across three subgroups (*P*=.04, Figure 6). We used pairwise multinomial logistic regression to determine factors independently associated with co-factor dominance (Figure 6C, Supplemental Table 4). C_{tail} editing sites, those with more mismatches in mooring and regulatory motif C, lower AU content in downstream sequence, and higher AU content in regulatory motif D were more likely co-dominant. Editing sites from small intestine and those with higher AU content of downstream sequence were more likely RBM47-dominant. Editing sites from liver and those with higher mismatches in regulatory motif B were more likely A1CF-dominant (Figure 6C).

Human mRNA targets

Finally, we turned to an analysis of human C-to-U RNA editing targets for which this same panel of parameters was available (Table 2). Aside from *APOB* RNA, which is known to be edited in the small intestine (Chen et al. 1987; Powell et al. 1987), other targets have been identified in central or peripheral nervous tissue (Skuse et al. 1996; Mukhopadhyay et al. 2002; Meier et al. 2005; Schaefermeier and Heinze 2017). The human targets were categorized into low editing (*NF1, GLYRα2, GLYRα3*) and high editing (*APOB*, *TPH2B* exon3, *TPH2B* exon7) subgroups using 20% as cut-off. A composite score (maximum=6) was generated based on six parameters introduced in the mouse model with notable variance between the two subgroups including mismatches in mooring sequence, spacer length, location of the edited cytidine, and relative abundance of stem-loop bases (Table 2). High editing targets exhibited a significantly higher composite score (4.7 vs 2, P=.001) compared to low editing targets and the composite score significantly correlated with editing frequency in individual targets (r=0.95, P=.005). The canonical editing target ApoB (Chen et al. 1987; Powell et al. 1987) achieved a score of 5 (out of 6), reflecting the observation that one of the six parameters (AU% of regulatory motifs) in human *APOB* is non-preferential compared to the editing-promoting features identified in the mouse multivariable model.

DISCUSSION

The current study reflects our analysis of 177 C-to-U RNA editing sites from 119 target mRNAs, with the majority residing within the 3' untranslated region. Our multivariable model identified several key factors influencing editing frequency, including host tissue, base content of nucleotides surrounding the edited cytidine, number of mismatches in regulatory and mooring sequences, AU content of the regulatory sequence, overall secondary structure, location of the mooring sequence, and co-factor dominance. These factors, each exerting independent effects, together accounted for 84% of the variance in editing frequency. Our findings also showed that mismatches in the mooring and regulatory sequences, AU content of regulatory and downstream sequences, host tissue and secondary structure of target mRNA were associated with the pattern of co-factor dominance. Several aspects of these primary conclusions merit further discussion.

Previous studies investigating the key factors that regulate C-to-U mRNA editing were confined to *in vitro* studies and predicated on a single mRNA target (*ApoB*) (Backus and Smith 1991; Shah et al. 1991; Smith et al. 1991; Backus and Smith 1992; Hersberger and Innerarity 1998). With the expanded range of verified C-to-U RNA editing targets now available for interrogation, we revisited the original assumptions to understand more globally the determinants of C-to-U mRNA editing efficiency. In undertaking this analysis, we were reminded that the requirements for C-to-U mRNA editing *in vitro* often appear more stringent than *in vivo* (Backus and Smith 1991; Shah et al. 1991), which further emphasizes the importance of our findings. In addition, our approach included both *cis-acting* sequence- and folding-related predictions along with the role of *trans*-acting factors and took advantage of statistical modeling to adjust for confounding or modifier effects between these factors to identify their role in editing frequency.

We began with the assumptions established for *Apob* RNA editing which identified a 26 nucleotide segment encompassing the edited base, spacer, mooring sequence, and part of regulatory sequence as the minimal sequence competent for physiological editing *in vitro* and *in vivo* (Davies et al. 1989; Shah et al. 1991; Backus and Smith 1992). Those studies identified an 11-nucleotide mooring sequence as essential and sufficient for editosome assembly and site-specific C-to-U editing (Backus and Smith 1991; Shah et al. 1991; Backus and Smith 1992) and established optimal positioning of the mooring sequence relative to the edited base in *Apob* RNA (Backus and Smith 1992). The current work supports the key conclusions of this original mooring sequence model as applied to the entire range of C-to-U RNA editing targets. We observed that mismatches in either the mooring or regulatory sequences were independent factors governing editing frequency. By contrast, while mismatches in the spacer sequence also showed negative association with editing frequency, the impact of spacer mismatches were not retained in the final model, nor was the length of the spacer associated with editing frequency. Furthermore, we found mismatches in the regulatory sequence motif C to be more important than mismatches in motif B. These inconsistencies might conceivably reflect the context in which an RNA segment is studied (Backus and Smith 1992). For example, our analysis reflects physiological conditions in which naturally occurring mRNA targets are edited, while the aforementioned study used *in vitro* data based on varying lengths of *Apob* mRNA embedded within different mRNA contexts (*Apoe* RNA) (Backus and Smith 1992).

In addition to the components of mooring sequence model, we examined variations in the base content in different segments/motifs as well as among individual nucleotides surrounding the edited cytidine. As expected, we found that sequences flanking the edited cytidine exhibited high AU content. We further observed a similarly high AU content in the flanking sequences of a range of proposed APOBEC-mediated DNA mutation targets in human cancer tissues and cell lines (Alexandrov et al. 2013; Petljak et al. 2019), especially in targets with dC/dT change (Nik-

Zainal et al. 2012). This observation implies that APOBEC-mediated DNA and RNA editing frequency may each be functionally modified by AU enrichment in the flanking sequences surrounding modifiable bases. The base content in individual nucleotides surrounding the edited cytidine also exerted significant impact on editing frequency, particularly in a 10-nucleotide segment spanning the edited cytidine (Supplemental Table 1), accounting for 25% of the variance in editing frequency independent of the mooring sequence model. Our findings regarding individual nucleotides surrounding the edited cytidine are consistent with findings for both DNA and RNA editing targets, particularly in the setting of cancers (Backus and Smith 1992; Conticello 2012; Roberts et al. 2013; Saraconi et al. 2014; Gao et al. 2018; Arbab et al. 2020). Recent work examining the sequence-editing relationship of a large *in vitro* library of DNA targets edited by different synthetic cytidine base editor (CBE)s (Arbab et al. 2020) showed that the base content of a 6-nucleotide window spanning the edited cytidine explained 23-57% of the editing variance, in particular one or two nucleotides immediately 5' of the edited nucleotide. That study also demonstrated that occurrence of T and C nucleotides at the position -1 increased, while a G nucleotide at that position decreased editing frequency (Arbab et al. 2020). However, in contrast to our findings, the presence of A at position -1 had either a negative or null effect on DNA editing activity (Arbab et al. 2020). This latter finding is consistent with the lower AU content observed in nucleotides adjacent to the edited cytidine in Apobec-1 DNA targets compared to the AU content in RNA targets. Our findings assign a greater importance of adjacent nucleotides in RNA editing frequency, similar to earlier reports that the five bases immediately 5' of the edited cytidine in *Apob* mRNA exert a greater impact on editing activity compared to nucleotides further upstream of this segment (Backus and Smith 1991; Shah et al. 1991; Backus and Smith 1992). G/C fraction of a 6-nucleotide window spanning the edited cytidine in DNA targets is associated with editing activity of the synthetic CBEs (Arbab et al. 2020). Although we found significant associations of RNA editing with G/C fraction in segments surrounding the edited cytidine in univariate analyses, these associations

were not retained in the final model. In contrast, the AU content of regulatory sequence motif B remained as an independent factor determining editing frequency in the final model.

The conserved 26-nucleotide sequence around the edited C forms a stem-loop secondary structure, where the editing site is in an octa-loop (Richardson et al. 1998) as predicted for the 55-nucleotide sequence of *ApoB* mRNA (Shah et al. 1991). This stem-loop structure is predicted to play an important role in recognition of the editing site by the editing factors (Bostrom et al. 1989; Davies et al. 1989; Driscoll et al. 1989; Chen et al. 1990). Mutations resulting in loss of base pairing in peripheral parts of the stem did not impact the editing frequency (Shah et al. 1991). Editing sites with the cytidine located in central parts (e.g. loop) exhibited higher editing frequencies than those with the edited cytidine located in peripheral parts (e.g. tail) and it is worth noting that the computer-based stem-loop structure was independently confirmed by NMR studies of a 31-nucleotide human *ApoB* mRNA (Maris et al. 2005). Those studies demonstrated that the location of the mooring sequence in the *ApoB* mRNA secondary structure plays a critical role in the RNA recognition by A1CF (Maris et al. 2005). In line with those findings, the current findings emphasize that the location of the mooring sequence in secondary structure of the target mRNA exerts significant independent impact on editing frequency. These predictions were confirmed in crystal structure studies of the carboxyl-terminal domain of APOBEC-1 and its interaction with cofactors and substrate RNA (Wolfe et al. 2020). Our conclusions regarding murine C-to-U editing frequency, such as mooring sequence, base content, and secondary structure appear consistent with a similar regulatory role among the smaller number of verified human targets. That being said, further study and expanded understanding of the range of C-to-U editing targets in human tissues will be needed as recently suggested (Destefanis et al. 2020), analogous to that for A-to-I editing (Bahn et al. 2012; Bazak et al. 2014).

We recognize that other factors likely contribute to the variance in RNA editing frequency not covered by our model. We did not consider the role of naturally occurring variants in APOBEC1, for example, which may be a relevant consideration since mutations in APOBEC family genes were shown to modify the editing activity of related hybrid DNA cytosine base editors (Arbab et al. 2020). Furthermore, genetic variants of *APOBEC1* in humans were associated with altered frequency of *GlyR* editing (Kankowski et al. 2017). Other factors not included in our approach included entropy-related features, tertiary structure of the mRNA target and other regulatory co-factors. Another limitation in the tissue-specific designation used to categorize editing frequency is that cell specific features of editing frequency may have been overlooked. For example, small intestinal and liver preparations are likely a blend of cell types (MacParland et al. 2018; Elmentaite et al. 2020) and tumor tissues are highly heterogeneous in cellular composition (Barker et al. 2009). The current findings provide a platform for future approaches to resolve these questions.

MATERIALS AND METHODS

Search strategy

A comprehensive literature review from 1987 (when *ApoB* RNA editing was first reported (Chen et al. 1987; Powell et al. 1987)) to November 2020, using studies published in English reporting C-to-U mRNA editing frequencies of individual or transcriptome-wide target genes. Databases searched included Medline, Scopus, Web of Science, Google Scholar, and ProQuest (for thesis). The references of full texts retrieved were also scrutinized for additional papers not indexed in the initial search.

Study selection

Primary records (N=528) were screened for relevance and *in vivo* studies reporting editing frequencies of individual or transcriptome-wide APOBEC1-dependent C-to-U mRNA targets selected, using a threshold of 10% editing frequency. For analyses based on RNA sequence information, only targets with available sequence information or chromosomal location for the edited cytidine were included. Exclusion criteria included: studies that reported C-to-U mRNA editing frequencies of target genes in other species, studies reporting editing frequencies of target genes in animal models overexpressing APOBEC1, exclusively *in vitro* studies, and conference abstracts.

Human targets

We included studies reporting human C-to-U mRNA targets (Chen et al. 1987; Powell et al. 1987; Skuse et al. 1996; Mukhopadhyay et al. 2002; Grohmann et al. 2010; Schaefermeier and Heinze 2017). We also included work describing APOBEC1-mediated mutagenesis in human breast cancer (Nik-Zainal et al. 2012).

Data extraction

Two reviewers (SS and VB) conducted the extraction process independently and discrepancies were addressed upon consensus and input from a third reviewer (NOD). The parameters were categorized as follows: *General parameters*: Gene name (RNA target), chromosomal and strand location of the edited cytidine, tissue site, editing frequency determined by RNA-seq or Sanger sequencing as illustrated for *ApoB* (Figure 1A). Editing frequency was highly correlated by both approaches (r=0.8 P<0.0001), and where both methodologies were available we used RNA-seq. We also defined relative dominance of editing co-factors (A1CF-dominant, RBM47-dominant, or co-dominant), relative mRNA expression (edited gene vs unedited gene) by RNA-seq or quantitative RT-PCR, and abundance of corresponding protein (edited gene vs unedited gene) by western blotting or proteomic comparison. Co-factor dominancy was determined based on the relative contribution of each co-factor to editing frequency. In each editing site, editing frequencies in mouse tissues deficient in *A1cf* or *Rbm47* were compared to that of wild-type mice. The relative contribution of each co-factor was calculated by subtracting the editing frequency for each target in *A1cf* or *Rbm47* knockout tissue from the total editing frequency in wild-type control. Editing sites with <20% difference between contributions of RBM47 and A1CF were considered co-dominant. Sites with ≥20% difference were considered either RBM47- or A1CF-dominant, depending on the co-factor with higher contribution (Blanc et al. 2019).

Sequence-related parameters: A sequence spanning 10 nucleotides upstream and 30 nucleotides downstream of the edited cytidine was extracted for each C-to-U mRNA editing site. These sequences were extracted either directly from the full-text or using online UCSC Genome Browser on Mouse (NCBI37/mm9) and Human (Grch38/hg38) (https://genome.ucsc.edu/cgi-bin/hgGateway) . Using the mooring sequence model (Backus and Smith 1992), three *cis*-acting elements were considered for each site. These elements included 1) a 10-nucleotide segment immediately upstream of the edited cytidine as “regulatory sequence”; 2) a 10-nucleotide segment downstream of the edited cytidine with complete or partial consensus with the

canonical “mooring sequence” of *ApoB* mRNA; 3) the sequence between the edited cytidine and the 5’ end of the mooring sequence, referred to as “spacer”. We used an unbiased approach to identify potential mooring sequences by taking the nearest segment to the edited cytidine with lowest number of mismatch(es) compared to the canonical mooring sequence of *ApoB* RNA. For each of the three segments, we investigated the number of mismatches compared to the corresponding segment of *ApoB* gene (Blanc et al. 2014), as well as length of spacer, the abundance of A and U nucleotides (AU content) and the G to C abundance ratio (G/C fraction (Arbab et al. 2020)). We also calculated relative abundance of A, G, C, and U individually across a region 10 nucleotides upstream and 20 nucleotides downstream of the edited cytidine across all editing sites. For comparison, we examined the base content of a sequence spanning 10 nucleotides upstream and downstream of mutated deoxycytidine for over 6000 proposed C to X (T, A, and G) DNA mutation targets of APOBEC family in human breast cancer (Nik-Zainal et al. 2012) along with relative deoxynucleotide distribution in proximity to the edited site.

Secondary structure parameters: We used RNA-structure (Reuter and Mathews 2010) and Mfold (Zuker 2003) to determine the secondary structure of an RNA cassette consisting of regulatory sequence, edited cytidine, spacer, and mooring sequence. Secondary structures similar to that of the cassette for *ApoB* chr12: 8014860 consisting of one loop and stem (with or without unassigned nucleotides with ≤4 unpaired bases inside the stem) as the main stem-loop with or without free tail(s) in one or both ends of the stem were considered as canonical. Two other types of secondary structure were considered as non-canonical structures (Figure 1B), with ≥2 loops located either at ends of the stem or inside the stem. Loops inside the stem were circular open structures with ≥5 unpaired bases. Editing sites with canonical structure were further categorized into three subgroups based on location of the edited cytidine: specifically (C_{loop}), stem (C_{stem}), or tail (C_{tail}). In addition to overall secondary structure, we considered

location of the edited cytidine, location of mooring sequence, symmetry of the free tails, and proportion of the nucleotides in the target cassette that constitute the main stem-loop. This proportion is 1.0 in the case of *ApoB* chr12: 8014860 where all the bases are part of the main stem-loop structure. Symmetry was defined based on existence of free tails in both ends of the RNA strand.

Statistical methodology

Continuous variables are reported as means ± SD with relative proportions for binary and categorical variables. *T-test* and *ANOVA* tests were used to compare continuous parameters of interest between two or more than two groups, respectively. *Chi-squared* testing was used to compare binary or categorical variables among different groups. Pearson *r* testing was used to investigate correlation of two continuous variables. We used linear regression analyses to develop the final model of independent factors that correlate with editing frequency. We used the Hosmer and Lemeshow approach for model building (Hosmer Jr et al. 2013) to fit the multivariable regression model. In brief, we first used bivariate and/or simple regression analyses with P value of 0.2 as the cut-off point to screen the variables and detect primary candidates for the multivariable model. Subsequently, we fitted the primary multivariable model using candidate variables from the screening phase. A backward elimination method was employed to reach the final multivariable model. Parameters with P values <0.05 or those that added to the model fitness were retained. Next, the eliminated parameters were added back individually to the final model to determine their impact. Plausible interaction terms between final determinants were also checked. The final model was screened for collinearity. We used the same approach to develop a multinomial logistic regression model to identify factors that were independently associated with co-factor dominance in RNA editing sites. Squared R and pseudo squared R were used to estimate the proportion of variance in responder parameter that could be explained by multivariable linear regression and multinomial logistic regression models,

respectively. The same screening and retaining methods were used to investigate association of base content in a sequence 10 nucleotides upstream and 20 nucleotides downstream of the edited cytidine, with editing frequency. However, after determining the nucleotides that were retained in final regression model, a proxy parameter named “base content score” was calculated for each editing site based on the β coefficient values retrieved for individual nucleotides in the model. This parameter was used in the final model as representative variable for base content of the aforementioned sequence in each editing site.

ACKNOWLEDGMENTS This work was supported by grants from the National Institutes of Health grants DK-119437, DK-112378, Washington University Digestive Diseases Research Core Center P30 DK-52574 (to NOD)

REFERENCES

UCSC Genome Browser on Mouse (NCBI37/mm9; 2007) and Human (GRCh38/hg38; 2013) assemblies.

Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL et al. 2013. Signatures of mutational processes in human cancer. *Nature* **500**: 415-421.

Arbab M, Shen MW, Mok B, Wilson C, Matuszek Z, Cassa CA, Liu DR. 2020. Determinants of Base Editing Outcomes from Target Library Analysis and Machine Learning. *Cell* **182**: 463-480 e430.

Backus JW, Schock D, Smith HC. 1994. Only cytidines 5' of the apolipoprotein B mRNA mooring sequence are edited. *Biochim Biophys Acta* **1219**: 1-14.

Backus JW, Smith HC. 1991. Apolipoprotein B mRNA sequences 3' of the editing site are necessary and sufficient for editing and editosome assembly. *Nucleic Acids Res* **19**: 6781-6786.

-. 1992. Three distinct RNA sequence elements are required for efficient apolipoprotein B (apoB) RNA editing in vitro. *Nucleic Acids Res* **20**: 6007-6014.

Bahn JH, Lee JH, Li G, Greer C, Peng G, Xiao X. 2012. Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. *Genome Res* **22**: 142-150.

Barker N, Ridgway RA, van Es JH, van de Wetering M, Begthel H, van den Born M, Danenberg E, Clarke AR, Sansom OJ, Clevers H. 2009. Crypt stem cells as the cells-of-origin of intestinal cancer. *Nature* **457**: 608-611.

Bazak L, Haviv A, Barak M, Jacob-Hirsch J, Deng P, Zhang R, Isaacs FJ, Rechavi G, Li JB, Eisenberg E et al. 2014. A-to-I RNA editing occurs at over a hundred million genomic sites, located in a majority of human genes. *Genome Res* **24**: 365-376.

Blanc V, Henderson JO, Newberry EP, Kennedy S, Luo J, Davidson NO. 2005. Targeted deletion of the murine apobec-1 complementation factor (acf) gene results in embryonic lethality. *Molecular and cellular biology* **25**: 7260-7269.

Blanc V, Park E, Schaefer S, Miller M, Lin Y, Kennedy S, Billing AM, Ben Hamidane H, Graumann J, Mortazavi A et al. 2014. Genome-wide identification and functional analysis of Apobec-1-mediated C-to-U RNA editing in mouse small intestine and liver. *Genome Biol* **15**: R79.

Blanc V, Xie Y, Kennedy S, Riordan JD, Rubin DC, Madison BB, Mills JC, Nadeau JH, Davidson NO. 2019. Apobec1 complementation factor (A1CF) and RBM47 interact in tissue-specific regulation of C to U RNA editing in mouse intestine and liver. *RNA* **25**: 70-81.

Bostrom K, Lauer SJ, Poksay KS, Garcia Z, Taylor JM, Innerarity TL. 1989. Apolipoprotein B48 RNA editing in chimeric apolipoprotein EB mRNA. *J Biol Chem* **264**: 15701-15708.

Chen SH, Habib G, Yang CY, Gu ZW, Lee BR, Weng SA, Silberman SR, Cai SJ, Deslypere JP, Rosseneu M et al. 1987. Apolipoprotein B-48 is the product of a messenger RNA with an organ-specific in-frame stop codon. *Science* **238**: 363-366.

Chen SH, Li XX, Liao WS, Wu JH, Chan L. 1990. RNA editing of apolipoprotein B mRNA. Sequence specificity determined by in vitro coupled transcription editing. *J Biol Chem* **265**: 6811-6816.

Conticello SG. 2012. Creative deaminases, self-inflicted damage, and genome evolution. *Annals of the New York Academy of Sciences* **1267**: 79-85.

Davies MS, Wallis SC, Driscoll DM, Wynne JK, Williams GW, Powell LM, Scott J. 1989. Sequence requirements for apolipoprotein B RNA editing in transfected rat hepatoma cells. *J Biol Chem* **264**: 13395-13398.

Destefanis E, Avsar G, Groza P, Romitelli A, Torrini S, Pir P, Conticello SG, Aguilo F, Dassi E. 2020. A mark of disease: how mRNA modifications shape genetic and acquired pathologies. *RNA*.

Driscoll DM, Wynne JK, Wallis SC, Scott J. 1989. An in vitro system for the editing of apolipoprotein B mRNA. *Cell* **58**: 519-525.

Elmentaite R, Ross ADB, Roberts K, James KR, Ortmann D, Gomes T, Nayak K, Tuck L, Pritchard S, Bayraktar OA et al. 2020. Single-Cell Sequencing of Developing Human Gut Reveals Transcriptional Links to Childhood Crohn's Disease. *Dev Cell*.

Fossat N, Tourle K, Radziewic T, Barratt K, Liebhold D, Studdert JB, Power M, Jones V, Loebel DA, Tam PP. 2014. C to U RNA editing mediated by APOBEC1 requires RNA-binding protein RBM47. *EMBO Rep* **15**: 903-910.

Gao J, Choudhry H, Cao W. 2018. Apolipoprotein B mRNA editing enzyme catalytic polypeptide-like family genes activation and regulation during tumorigenesis. *Cancer science* **109**: 2375-2382.

Giannoni F, Bonen DK, Funahashi T, Hadjiagapiou C, Burant CF, Davidson NO. 1994. Complementation of apolipoprotein B mRNA editing by human liver accompanied by secretion of apolipoprotein B48. *J Biol Chem* **269**: 5932-5936.

Grohmann M, Hammer P, Walther M, Paulmann N, Buttner A, Eisenmenger W, Baghai TC, Schule C, Rupprecht R, Bader M et al. 2010. Alternative splicing and extensive RNA editing of human TPH2 transcripts. *PloS one* **5**: e8956.

Gu T, Buaas FW, Simons AK, Ackert-Bicknell CL, Braun RE, Hibbs MA. 2012. Canonical A-to-I and C-to-U RNA editing is enriched at 3'UTRs and microRNA target sites in multiple mouse tissues. *PLoS One* **7**: e33720.

Harris RS, Bishop KN, Sheehy AM, Craig HM, Petersen-Mahrt SK, Watt IN, Neuberger MS, Malim MH. 2003. DNA deamination mediates innate immunity to retroviral infection. *Cell* **113**: 803-809.

Hersberger M, Innerarity TL. 1998. Two efficiency elements flanking the editing site of cytidine 6666 in the apolipoprotein B mRNA support mooring-dependent editing. *J Biol Chem* **273**: 9435-9442.

Hirano K, Young SG, Farese RV, Jr., Ng J, Sande E, Warburton C, Powell-Braxton LM, Davidson NO. 1996. Targeted disruption of the mouse apobec-1 gene abolishes apolipoprotein B mRNA editing and eliminates apolipoprotein B48. *J Biol Chem* **271**: 9887-9890.

Hosmer Jr DW, Lemeshow S, Sturdivant RX. 2013. *Applied logistic regression*. John Wiley & Sons.

Hospattankar AV, Higuchi K, Law SW, Meglin N, Brewer HB, Jr. 1987. Identification of a novel in-frame translational stop codon in human intestine apoB mRNA. *Biochem Biophys Res Commun* **148**: 279-285.

Kanata E, Llorens F, Dafou D, Dimitriadis A, Thune K, Xanthopoulos K, Bekas N, Espinosa JC, Schmitz M, Marin-Moreno A et al. 2019. RNA editing alterations define manifestation of prion diseases. *Proc Natl Acad Sci U S A* **116**: 19727-19735.

Kankowski S, Forstera B, Winkelmann A, Knauff P, Wanker EE, You XA, Semtner M, Hetsch F, Meier JC. 2017. A Novel RNA Editing Sensor Tool and a Specific Agonist Determine Neuronal Protein Expression of RNA-Edited Glycine Receptors and Identify a Genomic APOBEC1 Dimorphism as a New Genetic Risk Factor of Epilepsy. *Front Mol Neurosci* **10**: 439.

Lellek H, Kirsten R, Diehl I, Apostel F, Buck F, Greeve J. 2000. Purification and molecular cloning of a novel essential component of the apolipoprotein B mRNA editing enzyme-complex. *J Biol Chem* **275**: 19848-19856.

MacParland SA, Liu JC, Ma XZ, Innes BT, Bartczak AM, Gage BK, Manuel J, Khuu N, Echeverri J, Linares I et al. 2018. Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. *Nat Commun* **9**: 4383.

Maris C, Masse J, Chester A, Navaratnam N, Allain FH. 2005. NMR structure of the apoB mRNA stem-loop and its interaction with the C to U editing APOBEC1 complementary factor. *RNA* **11**: 173-186.

Mehta A, Kinter MT, Sherman NE, Driscoll DM. 2000. Molecular cloning of apobec-1 complementation factor, a novel RNA-binding protein involved in the editing of apolipoprotein B mRNA. *Mol Cell Biol* **20**: 1846-1854.

Meier JC, Henneberger C, Melnick I, Racca C, Harvey RJ, Heinemann U, Schmieden V, Grantyn R. 2005. RNA editing produces glycine receptor alpha3(P185L), resulting in high agonist potency. *Nat Neurosci* **8**: 736-744.

Mukhopadhyay D, Anant S, Lee RM, Kennedy S, Viskochil D, Davidson NO. 2002. C-->U editing of neurofibromatosis 1 mRNA occurs in tumors that express both the type II transcript and apobec-1, the catalytic subunit of the apolipoprotein B mRNA-editing enzyme. *Am J Hum Genet* **70**: 38-50.

Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, Jones D, Hinton J, Marshall J, Stebbings LA et al. 2012. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**: 979-993.

Petljak M, Alexandrov LB, Brammeld JS, Price S, Wedge DC, Grossmann S, Dawson KJ, Ju YS, Iorio F, Tubio JMC et al. 2019. Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis. *Cell* **176**: 1282-1294 e1220.

Powell LM, Wallis SC, Pease RJ, Edwards YH, Knott TJ, Scott J. 1987. A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine. *Cell* **50**: 831-840.

Rayon-Estrada V, Harjanto D, Hamilton CE, Berchiche YA, Gantman EC, Sakmar TP, Bulloch K, Gagnidze K, Harroch S, McEwen BS et al. 2017. Epitranscriptomic profiling across cell types reveals associations between APOBEC1-mediated RNA editing, gene expression outcomes, and cellular function. *Proc Natl Acad Sci U S A* **114**: 13296-13301.

Reuter JS, Mathews DH. 2010. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* **11**: 129.

Richardson N, Navaratnam N, Scott J. 1998. Secondary structure for the apolipoprotein B mRNA editing site. Au-binding proteins interact with a stem loop. *J Biol Chem* **273**: 31707-31717.

Roberts SA, Lawrence MS, Klimczak LJ, Grimm SA, Fargo D, Stojanov P, Kiezun A, Kryukov GV, Carter SL, Saksena G et al. 2013. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat Genet* **45**: 970-976.

Rosenberg BR, Hamilton CE, Mwangi MM, Dewell S, Papavasiliou FN. 2011. Transcriptome-wide sequencing reveals numerous APOBEC1 mRNA-editing targets in transcript 3' UTRs. *Nat Struct Mol Biol* **18**: 230-236.

Saraconi G, Severi F, Sala C, Mattiuz G, Conticello SG. 2014. The RNA editing enzyme APOBEC1 induces somatic mutations and a compatible mutational signature is present in esophageal adenocarcinomas. *Genome Biol* **15**: 417.

Schaefermeier P, Heinze S. 2017. Hippocampal Characteristics and Invariant Sequence Elements Distribution of GLRA2 and GLRA3 C-to-U Editing. *Mol Syndromol* **8**: 85-92.

Shah RR, Knott TJ, Legros JE, Navaratnam N, Greeve JC, Scott J. 1991. Sequence requirements for the editing of apolipoprotein B mRNA. *J Biol Chem* **266**: 16301-16304.

Skuse GR, Cappione AJ, Sowden M, Metheny LJ, Smith HC. 1996. The neurofibromatosis type I messenger RNA undergoes base-modification RNA editing. *Nucleic Acids Res* **24**: 478-485.

Smith HC, Kuo SR, Backus JW, Harris SG, Sparks CE, Sparks JD. 1991. In vitro apolipoprotein B mRNA editing: identification of a 27S editing complex. *Proc Natl Acad Sci U S A* **88**: 1489-1493.

Snyder EM, McCarty C, Mehalow A, Svenson KL, Murray SA, Korstanje R, Braun RE. 2017. APOBEC1 complementation factor (A1CF) is dispensable for C-to-U RNA editing in vivo. *RNA* **23**: 457-465.

Sowden M, Hamm JK, Spinelli S, Smith HC. 1996. Determinants involved in regulating the proportion of edited apolipoprotein B RNAs. *RNA* **2**: 274-288.

Teng B, Burant CF, Davidson NO. 1993. Molecular cloning of an apolipoprotein B messenger RNA editing protein. *Science* **260**: 1816-1819.

Wolfe AD, Arnold DB, Chen XS. 2019. Comparison of RNA Editing Activity of APOBEC1-A1CF and APOBEC1-RBM47 Complexes Reconstituted in HEK293T Cells. *J Mol Biol* **431**: 1506-1517.

Wolfe AD, Li S, Goedderz C, Chen XS. 2020. The structure of APOBEC1 and insights into its RNA and DNA substrate selectivity. *NAR Cancer* **2**: zcaa027.

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

Table 1. Multivariable linear regression model for determinant factors of editing frequency in mouse APOBEC1-dependent C-to-U mRNA editing sites.

Determinant of editing frequency	Subgroup	ß (95% CI)	*P* value
Model without co-factor group N=103; R^2= 0.84; *P*<.001			
Base content score	per unit increments	1.00 [0.83, 1.17]	<0.001
Count of mismatches in mooring sequence	per unit increments	-5.89 [-7.48, -4.31]	<.001
Count of mismatches in regulatory sequence motif D (whole sequence)	per unit increments	-2.00 [-3.58, -0.43]	.01
AU content of regulatory sequence motif B	per 10% increments	-2.41 [-4.38, -0.45]	.02
Overall secondary structure	C_{loop}	Reference	
	C_{stem}	1.20 [-5.07, 7.47]	.7
	C_{tail}	-12.19 [-20.80, -3.58]	.006
	Non-canonical	-10.67 [-20.92, -0.43]	0.04
Location of mooring sequence	Stem-loop	Reference	
	Other	-11.56 [-17.35, -5.77]	<.001
After adding co-factor group to the model N=72; R^2= 0.84; *P*<.001			
Co-factor group	RBM47 dominant	Reference	
	Co-dominant	-12.30 [-20.63, -3.97]	.005
	A1CF dominant	11.54 [-0.64, 23.72]	.07

ß: represents average change (%) in the editing frequency compared to the reference group

CI: confidence interval

Table 2: Characteristics of human C-to-U mRNA editing targets

Parameter	Low editing			High editing		
	NF1	GLYCRA3	GLYCRA2	TPH2B	TPH2B	APOB
Editing location	C2914	C554	C575	C385 (exon3)	C830 (exon7)	C6666
Tissue	neural sheath / CNS tumor	hippocampus	hippocampus	amygdala	amygdala	small intestine
Editing frequency %)	10	10	17	89	98	>95
Mismatches in regulatory motif A	1	3	3	2	3	0
Mismatches in regulatory motif B	2	4	5	4	5	0
Mismatches in regulatory motif C	4	4	4	4	4	0
Mismatches in regulatory motif D	6	8	9	8	9	0
AU content (%) in regulatory motif A	100	33	33	100	0	100
AU content (%) in regulatory motif B	100	60	20	100	20	80
AU content (%) in regulatory motif C*	60	40	60	40	40	100
AU content (%) in regulatory motif D	80	50	40	70	30	90
Spacer length*	6	2	2	0	3	4
Spacer AU content (%)	67	0	0		33	100
Mismatches in spacer	2	2	2		2	0
Mismatches in mooring*	3	4	2	1	5	0
AU content (%) of 3 downstream bases*	67	33	33	100	33	100
AU content (%) of 20 downstream bases	60	60	70	55	35	85
Overall secondary structure	canonical	canonical	canonical	canonical	canonical	canonical
Location of edited C*	loop	tail	tail	stem	loop	loop
Location of mooring sequence	stem-loop	stem-loop	stem-loop	stem-loop	stem-loop	stem-loop
Ratio of stem-loop bases*	0.46	0.375	0.5	0.45	0.92	0.96
Free tail orientation	symmetric	symmetric	asymmetric	symmetric	asymmetric	asymmetric
Composite score	2	2	2	5	4	5

CNS: central nervous system

* these items were used to calculate the composite score (total score = 6) as follows:
AU content (%) in regulatory motif C: < 50%: 1, ≥ 50%: 0
spacer length: ≤ 4: 1, > 4: 0
mismatches in mooring: < 3: 1, ≥ 3: 0
AU content (%) of 3 downstream bases: > 50%: 1, ≤ 50%: 0
location of edited C in secondary structure: stem-loop: 1, tail: 0
ratio of stem-loop bases: > 50%: 1, ≤ 50%: 0

FIGURE LEGENDS

Figure 1. Characteristics of murine APOBEC1-mediated C-to-U mRNA editing sites. A: schematic presentation of mRNA target, chromosomal editing location, and editing sites considered. Each mRNA target could be edited at one or more chromosomal location(s) (blue boxes). Each editing location could be edited in one or more tissues giving rise to one or more editing site(s) per location (green boxes). Editing site(s) of each mRNA target are the sum of editing sites from all editing locations reported for that target. B: examples of canonical (*ApoB* chr12: 8014860, top) and two types of non-canonical (*Kctd12* chr14: 103379573 and *Dcn* chr10: 96980535) secondary structures. C: distribution of number of chromosomal editing location(s), or targeted cytidine(s), per mRNA target. D: distribution of number of total editing sites per mRNA target considering all chromosomal location(s) edited at different tissue(s). E: distribution of location of editing sites within gene structure.

Figure 2. Base content of sequences flanking modified cytidine in RNA editing and DNA mutation targets. A: base content of 10 nucleotides upstream and 20 nucleotides downstream of edited cytidine in mouse APOBEC1-mediated C-to-U mRNA editing targets. B: base content of 10 nucleotides upstream and 10 nucleotides downstream of mutated cytidine in proposed human APOBEC-mediated DNA mutation targets in patients with breast cancer. C: comparison of AU base content (%) of nucleotides flanking modified cytidine in RNA editing targets and DNA mutation targets in mouse and human breast cancer patients, respectively.

Figure 3. Characteristics of regulatory-spacer-mooring cassette and base content of individual nucleotides flanking edited cytidine in association with editing frequency. A: schematic illustration of regulatory-spacer-mooring cassette. Four motifs were defined for

regulatory sequence: motif A for nucleotides -1 to -3; motif B for nucleotides -1 to -5; motif C for nucleotides -6 to -10; motif D representative of the whole sequence. B: association of the mismatches in motif D of regulatory sequence with editing frequency. C: association between the AU content (%) of regulatory sequence (motif B) and editing frequency. D: association of the mismatches in spacer (nucleotides +1 to +4 downstream of the edited cytidine) with editing frequency. E: association of the mismatches in mooring sequence with editing frequency. F: heatmap plot illustrating the association between base content of 30 nucleotides flanking the edited cytidine with editing frequency. Red color density in each cell represents the beta coefficient value of corresponding base in the multivariable linear regression model fit including that nucleotide. The asteriska refer to the nucleotides that were retained in the final model. Mismatches in regulatory, spacer, and mooring sequences were determined in comparison to the corresponding sequences in *ApoB* mRNA (as reference). *r*: Pearson correlation coefficient.

Figure 4. Secondary structure-related features in association with editing frequency. A: distribution of different types of overall secondary structure in editing sites. C_{loop}, C_{stem}, C_{tail} are three subtypes of canonical secondary structure based on the location of the edited cytidine. B: association between type of secondary structure and editing frequency. C: distribution of the mooring sequence location in editing sites. “Other” refers to mooring sequences located in tail or stem/loop and not part of the main stem-loop structure. D: association of mooring sequence location with editing frequency. E: association between ratio of main stem-loop bases to total bases count and editing frequency. F: association of the 5’ free tail length with editing frequency. * $P<.05$; ** $P<.001$. *r*: Pearson correlation coefficient.

Figure 5. **Dominance and tissue-specific cofactor patterns among editing sites.** A: distribution of dominant co-factor in editosomes of editing sites. B: association of dominant co-factor with editing frequency. C: distribution of number of editing tissue(s) per mRNA target. D: tissue distribution of editing sites. E: average editing frequency of editing sites edited at different tissues. SI, small intestine.

Figure 6. **Co-factor pattern and tissue-specific role in murine C-to-U mRNA editing sites.** A: distribution of editing tissue across subgroups of editing sites with different dominant co-factor patterns. B: location of edited cytidine in secondary structure of editing sites with different dominant co-factor patterns. C: schematic presentation of factors that correlate with dominant co-factor pattern in editing sites. This graph is based on the findings derived from pairwise multinomial logistic regression models.

SUPPLEMENTAL FIGURE LEGENDS

Supplemental Figure 1. Chromosomal distribution of murine APOBEC1-mediated C-to-U mRNA editing sites. The black curve corresponds to left Y-axis and represents average editing frequencies of editing sites related to each chromosome. The blue curve corresponds to right Y axis and represents number of editing sites related to each chromosome.

Supplemental Figure 2. Association of editing frequency with characteristics of regulatory sequence in murine APOBEC1-mediated C-to-U mRNA editing sites. A-C**.** Association of editing frequency with number of mismatches and AU content (%). D-F Association of editing frequency with different regulatory sequence motifs. Mismatches were determined in comparison to the same regulatory sequence motif in *ApoB* mRNA (as reference).

Supplemental Figure 3. Association of editing frequency with characteristics of downstream sequence in murine APOBEC1-mediated C-to-U mRNA editing sites. A. Association of editing frequency with spacer length. B. Association of editing frequency with spacer AU content (%). C-F. Association of editing frequency with and AU content of successive segments downstream of the edited cytidine.

Supplemental Figure 4. Association of editing frequency with secondary structure-related characteristics in C-to-U mRNA editing sites. A: distribution of edited cytidine location in secondary structure regardless of the overall secondary structure. B: association of editing frequency with edited cytidine location in secondary structure. C: distribution of free tail

orientation in editing sites. D: association of editing frequency with free tail orientation in editing sites. E: association of editing frequency with 3' free tail length. * *P*<.05; *** *P*<.0001. *r*: Pearson correlation coefficient.

Supplemental Figure 5. Association of secondary structure-related characteristics with dominant co-factor pattern in APOBEC1-mediated C-to-U mRNA editing sites. A. Distribution of mooring sequence location presented in the context of different dominant co-factor patterns. B. Distribution of free tail orientation in secondary structure among editing sites, presented in the context of different dominant co-factor patterns.

Supplemental table 1. Multivariable linear regression model for individual nucleotides surrounding edited cytosine (-10 to +20) in mouse APOBEC1-dependent C-to-U mRNA editing sites.

Location of nucleotide relative to edited C	Base preference	ß (95% CI)	*P* value
Nucleotide -8	GU	8.15 [3.0,13.3]	0.002
Nucleotide -7	C	12.7 [4.3, 21.0]	0.003
Nucleotide -6	G	7.1 [0.6, 13.7]	0.03
Nucleotide -5	U	5.2 [1.0, 9.5]	0.02
Nucleotide -2	AUC	13.5 [9.0, 17.9]	<0.001
Nucleotide -1	AU	15.9 [4.0, 27.9]	0.01
Nucleotide +1	AGU	19.5 [12.5, 26.6]	<0.001
Nucleotide +3	G	12.2 [7.4, 16.9]	<0.001
Nucleotide +4	G	15.9 [10.9, 21.0]	<0.001
Nucleotide +7	C	10.3 [1.5, 19.2]	0.02
Nucleotide +9	G	9.7 [1.4, 18.0]	0.02
Nucleotide +12	AUC	7.5 [1.0, 13.9]	0.02
Nucleotide +16	AC	6.6 [2.2, 11.0]	0.004
Nucleotide +17	AU	5.6 [0.5, 10.8]	0.03
Nucleotide +18	AU	6.6 [1.5, 11.8]	0.01
Nucleotide +19	AC	5.65 [1.3, 10.0]	0.01

ß: represents average change (%) in the editing frequency compared to the reference group (non-preferred group)
CI: confidence interval

Supplemental table 2. Descriptive data of regulatory-spacer-mooring cassette in mouse APOBEC1-dependent C-to-U mRNA editing sites.					
Parameter	N	Mean	SD	Min	Max
Sequence-related features					
Mismatches in regulatory (motif A)	177	1.72	0.94	0	3
Mismatches in regulatory (motif B)	177	3.35	1.12	0	5
Mismatches in regulatory (motif C)	177	3.78	0.99	0	5
Mismatches in regulatory (motif D)	177	7.12	1.76	0	10
AU content (%) of regulatory (motif A)	177	75.14	26.00	0	100
AU content (%) of regulatory (motif B)	177	73.44	22.10	0	100
AU content (%) of regulatory (motif C)	177	63.00	23.40	0	100
AU content (%) of regulatory (motif D)	177	68.25	18.40	10	100
Spacer length	177	5.08	3.67	0	20
Mismatches in spacer	152	2.54	1.09	0	4
AU content (%) of spacer	172	72.65	23.39	0	100
Mismatches in mooring	177	2.13	1.81	0	8
AU content (%) of downstream sequence +1 to +5	177	72.88	19.46	0	100
AU content (%) of downstream sequence +6 to +10	177	69.94	22.78	0	100
AU content (%) of downstream sequence +11 to +15	177	72.43	20.65	20	100
AU content (%) of downstream sequence +16 to +20	177	66.21	22.56	0	100
Secondary structure-related features					
Proportion of the bases that constitute main stem-loop	172	0.61	0.18	0.28	1
Length of 5’ free tail	172	4.25	3.93	0	15
Length of 3’ free tail	172	5.27	4.65	0	17
SD: standard deviation					

Supplemental table 3. Comparing three subgroups of mouse APOBEC1-dependent C-to-U mRNA editing sites based on co-factor dominance.

Parameter	RBM47-dominant			A1CF-dominant			Co-dominant			*P value*
	N	Mean	SD	N	Mean	SD	N	Mean	SD	
Mismatches in regulatory (motif A)	60	1.48	0.93	5	1.80	0.45	7	1.14	0.69	.4
Mismatches in regulatory (motif B)	60	3.05	1.13	5	3.60	0.55	7	3.00	0.82	.51
Mismatches in regulatory (motif C)	60	3.58	1.05	5	3.80	0.45	7	4.29	1.11	.1
Mismatches in regulatory (motif D)	60	6.63	1.90	5	7.40	0.55	7	7.29	1.50	.44
AU content (%) of regulatory (motif A)	60	82.22	18.88	5	80.00	18.26	7	85.71	17.82	.8
AU content (%) of regulatory (motif B)	60	76.33	16.67	5	84.00	16.73	7	82.86	17.99	.5
AU content (%) of regulatory (motif C)	60	62.67	22.84	5	72.00	17.89	7	62.86	21.38	.6
AU content (%) of regulatory (motif D)	60	69.50	14.89	5	78.00	13.04	7	72.86	12.54	.4
Spacer length	60	5.20	3.93	5	7.20	5.45	7	7.86	5.08	.2
Mismatches in spacer (in 4-base cassette)	40	2.43	1.20	4	2.75	1.50	6	3.83	0.41	.02
Mismatches in spacer (relative abundance (%))	60	61.81	30.89	5	61.67	36.13	7	82.14	37.40	.2
AU content (%) of spacer	60	77.30	17.83	5	72.08	18.14	7	71.37	15.24	.5
Mismatches in mooring	60	1.12	1.30	5	2.00	2.55	7	2.86	0.38	.004
AU content (%) of downstream sequence +1 to +5	60	77.33	14.94	5	80.00	20.00	7	71.43	15.74	.7
AU content (%) of downstream sequence +6 to +10	60	77.67	18.81	5	60.00	24.49	7	57.14	13.80	.01
AU content (%) of downstream sequence +11 to +15	60	80.33	15.40	5	72.00	17.89	7	65.71	15.12	0.06
AU content (%) of downstream sequence +16 to +20	60	70.33	20.00	5	72.00	10.95	7	77.14	17.99	.6
Proportion of the bases that constitute main stem-loop	60	0.62	0.18	5	0.71	0.10	7	0.59	0.21	.5
Length of 5' free tail	60	4.08	3.81	5	2.40	3.91	7	6.86	6.20	.3
Length of 3' free tail	60	5.35	4.84	5	6.00	2.55	7	5.00	5.66	.6

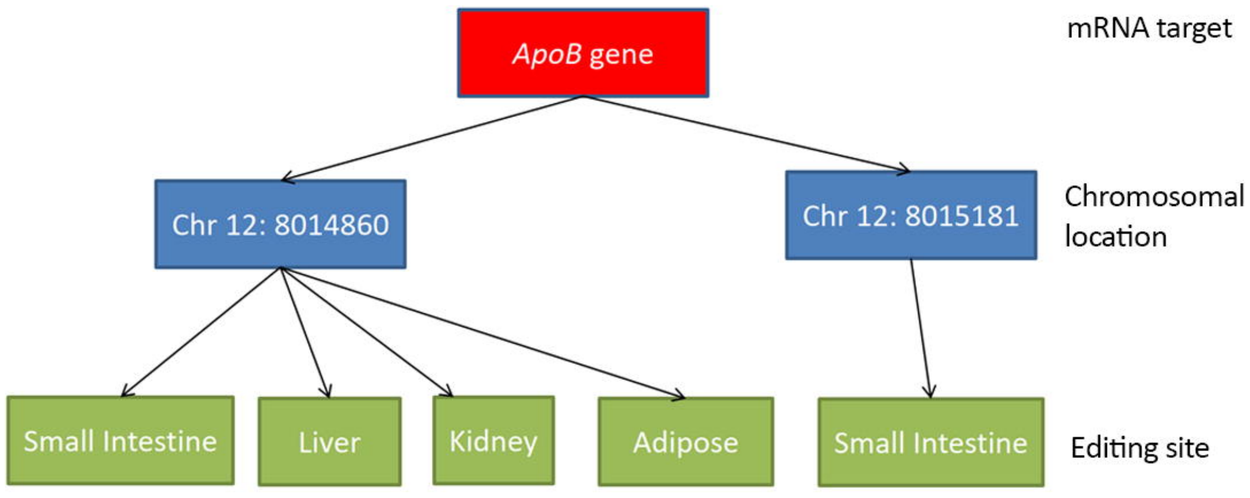
SD: standard deviation

Supplemental Table 4. Multinomial logistic regression model for determinant factors of co-factor dominancy in mouse APOBEC1-dependent C-to-U mRNA editing sites.

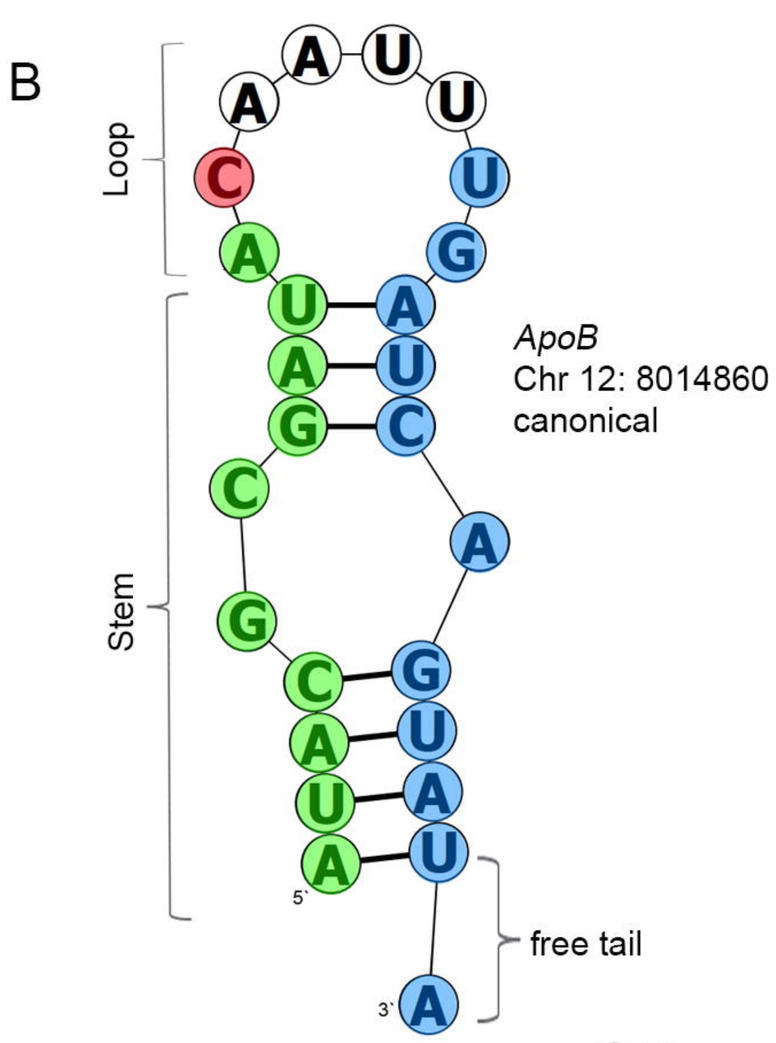
Determinant of co-factor dominancy	Subgroup	Coefficient (95% CI)	*P* value
A1CF-dominant vs RBM47-dominant			
Tissue	Small intestine	Reference	
	Liver	4.40 [0.34, 5.21]	.04
Location of edited cytosine	Loop	Reference	
	Stem	-3.88 [-8.31, 0.55]	0.08
	Tail	-19.13 [-25.82, -12.44]	<0.001
Mismatches in mooring sequence	per unit increments	0.30 [-0.97, 1.57]	0.6
Mismatches in regulatory sequence motif B	per unit increments	1.62 [0.063, 3.30]	.05
Mismatches in regulatory sequence motif C	per unit increments	0.12 [-0.83, 1.08]	.8
AU content (%) of regulatory sequence motif D	per unit increments	0.17 [-0.04, 0.39]	0.1
AU content (%) of downstream sequence +1 to +5	per unit increments	-0.02 [-0.09, 0.04]	0.5
AU content (%) of downstream sequence +6 to +10	per unit increments	-0.06 [-0.1, -0.02]	0.006
AU content (%) of downstream sequence +11 to +15	per unit increments	-0.06 [-0.18, 0.07]	0.4
Co-dominant vs RBM47-dominant			
Tissue	Small intestine	Reference	
	Liver	-1.73 [-6.00, 2.50]	0.4
Location of edited cytosine in secondary structure	C_{loop}	Reference	
	C_{stem}	1.70 [-2.11, 5.51]	0.4
	C_{tail}	3.70 [0.72, 6.67]	0.01
Mismatches in mooring sequence	per unit increments	0.66 [0.01, 1.33]	.05
Mismatches in regulatory sequence motif B	per unit increments	-2.32 [-3.86, -0.79]	.003
Mismatches in regulatory sequence motif C	per unit increments	3.16 [1.12, 5.21]	0.002
AU content (%) of regulatory sequence motif D	per unit increments	0.13 [0.02, 0.24]	0.02
AU content (%) of downstream sequence +1 to +5	per unit increments	-0.17 [-0.35, -0.01]	0.04
AU content (%) of downstream sequence +6 to +10	per unit increments	-0.10 [-0.28, 0.07]	0.25
AU content (%) of downstream sequence +11 to +15	per unit increments	-0.10 [-0.19, -0.01]	0.03
Co-dominant vs A1CF -dominant			
Tissue	Small intestine	Reference	
	Liver	-6.13 [-10.60, -0.31]	0.04
Location of edited cytosine in secondary structure	C_{loop}	Reference	
	C_{stem}	5.58 [0.06, 9.22]	0.05
	C_{tail}	22.83 [15.53, 30.12]	<0.001
Mismatches in mooring sequence	per unit increments	0.36 [-0.87, 1.59]	0.6
Mismatches in regulatory sequence motif B	per unit increments	-3.94 [-6.27, -1.61]	0.001
Mismatches in regulatory sequence motif C	per unit increments	3.04 [0.91, 5.16]	0.005
AU content (%) of regulatory sequence motif D	per unit increments	-0.04 [-0.29, 0.20]	0.72
AU content (%) of downstream sequence +1 to +5	per unit increments	-0.15 [-0.32, 0.02]	0.09
AU content (%) of downstream sequence +6 to +10	per unit increments	-0.04 [-0.22, 0.13]	0.62
AU content (%) of downstream sequence +11 to +15	per unit increments	-0.04 [-0.19, 0.11]	0.58
Model parameters: N=72; Pseudo R^2= 0.59; *P*<.001 CI: confidence interval			

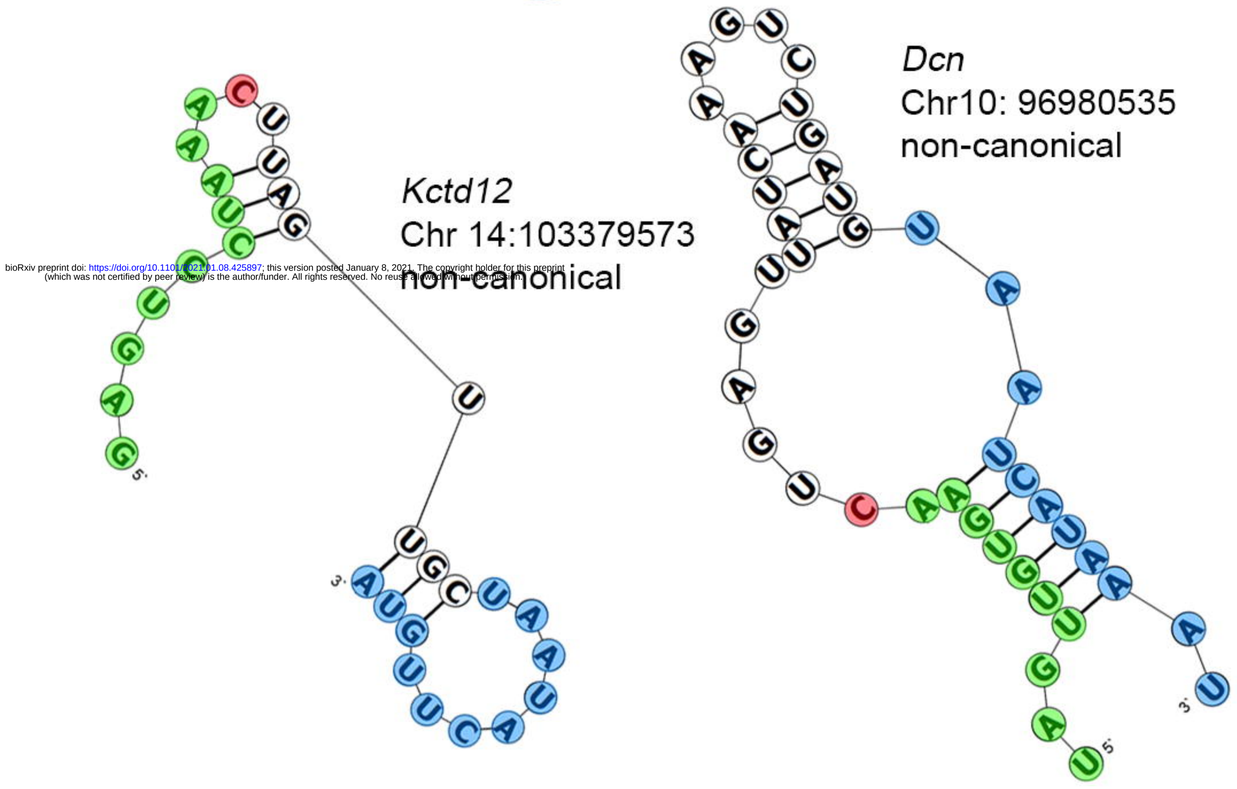
Figure 1

A



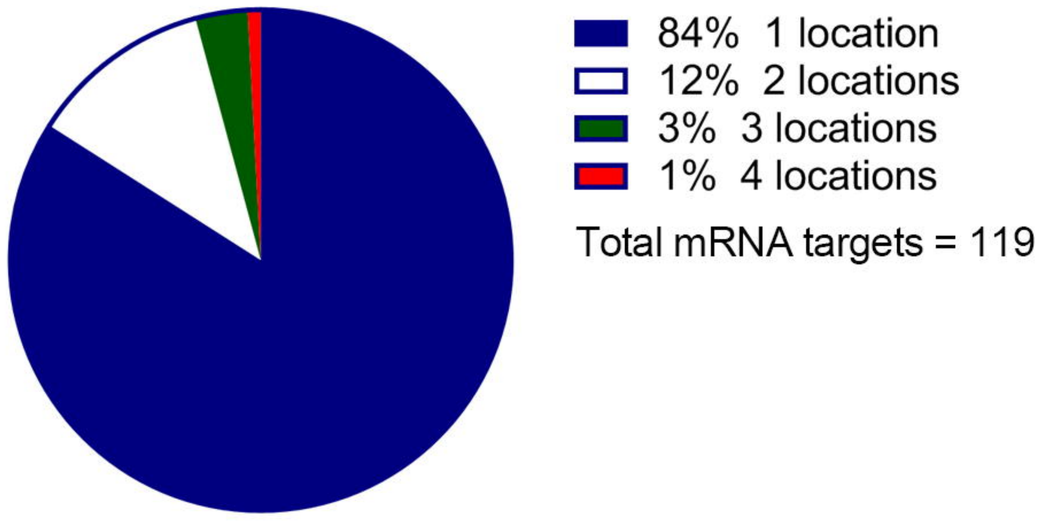
B





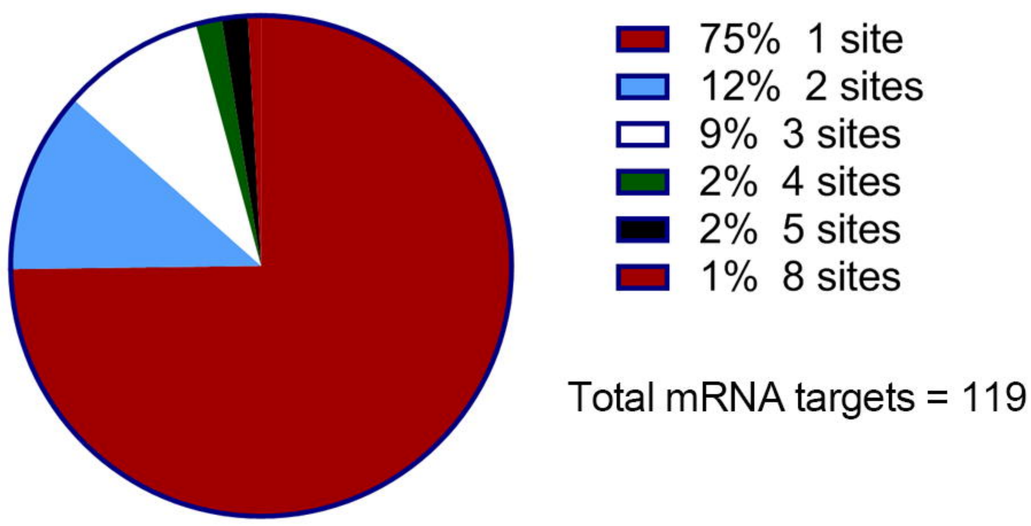
Edited cytidine
Regulatroy sequence
Mooring sequence

C



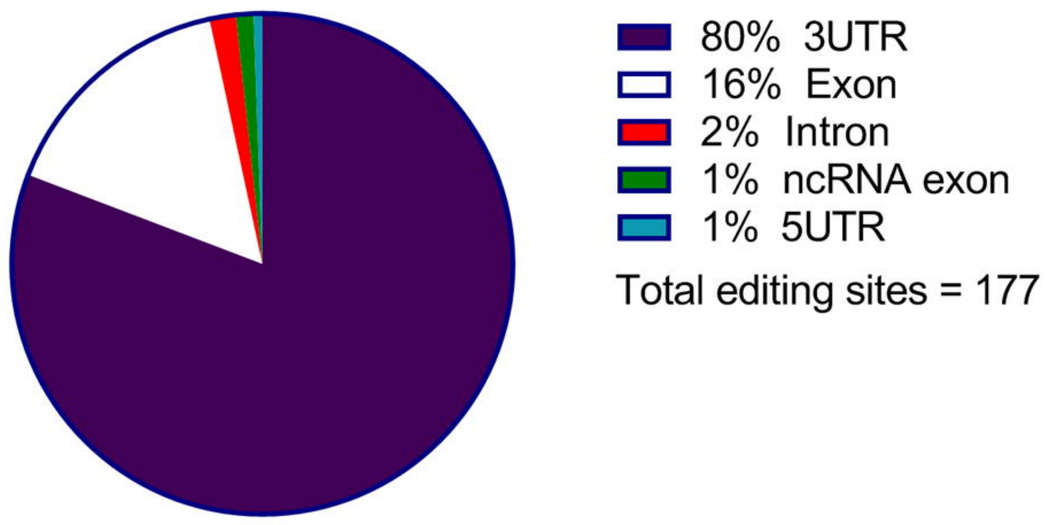
Distribution of C-to-U RNA editing locations per target

D



Distribution of C-to-U RNA editing sites per target

E



Distribution of location of C-to-U RNA editing sites in gene

Figure 2

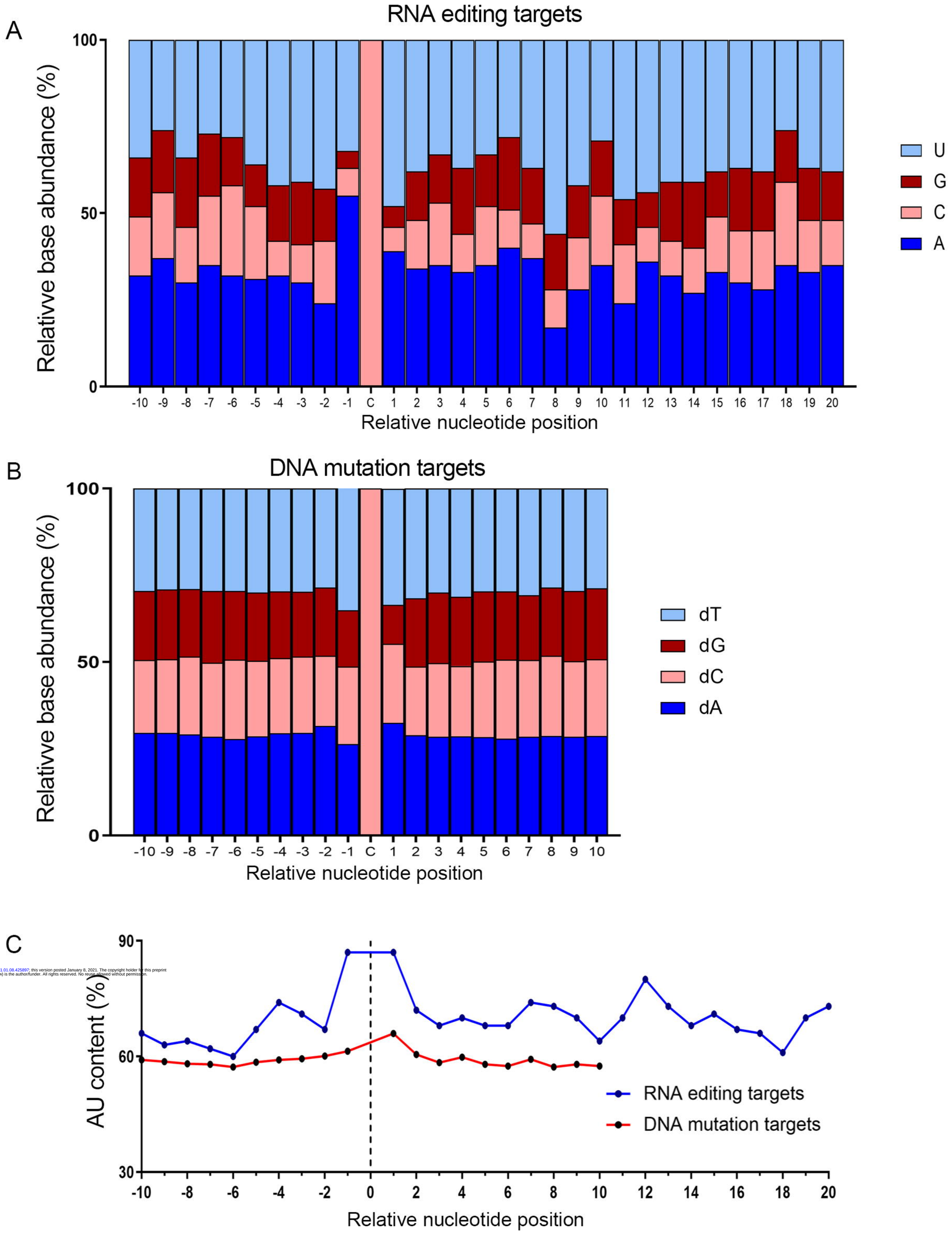
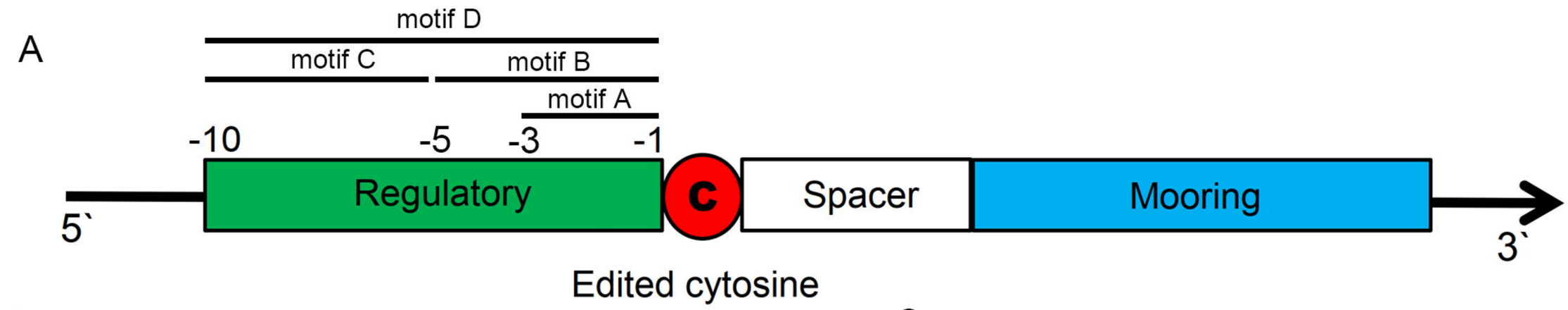
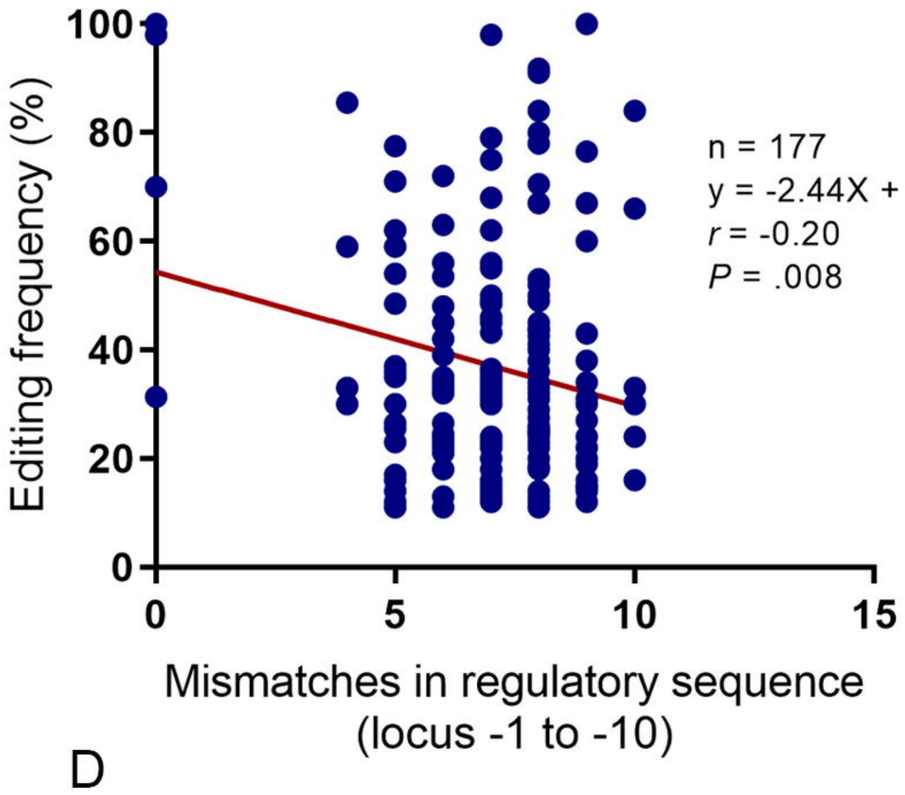
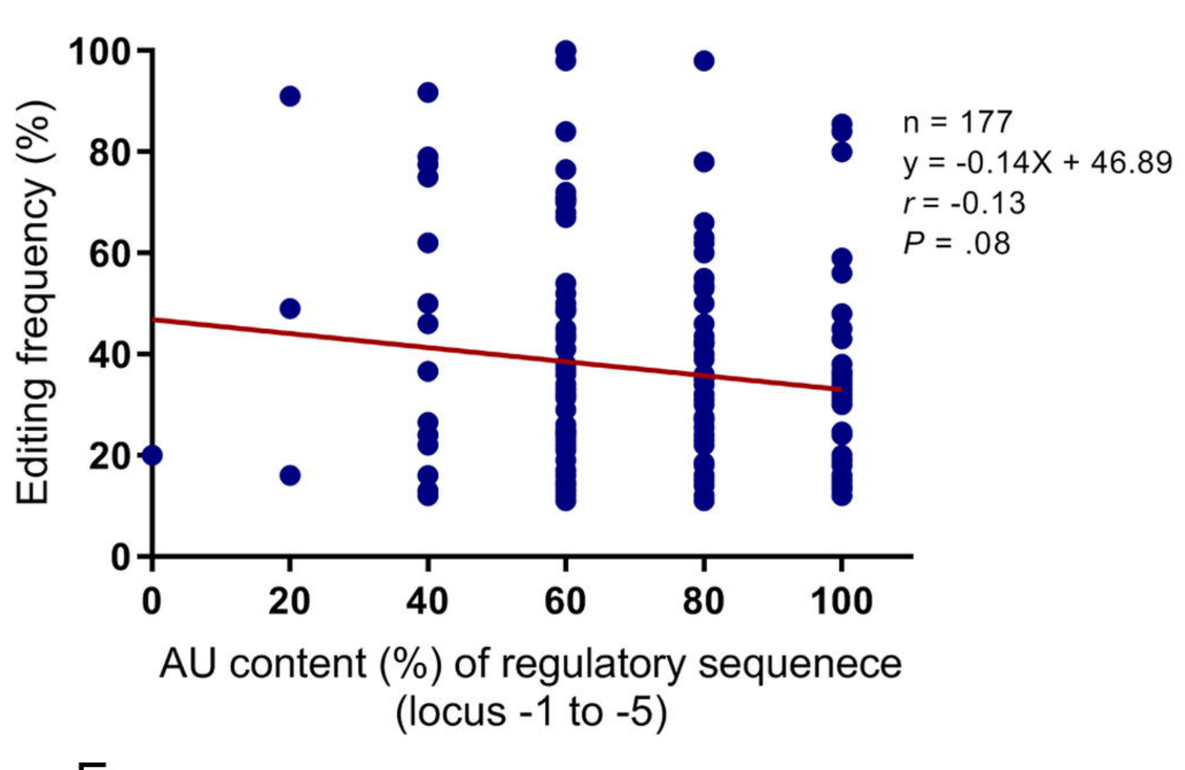
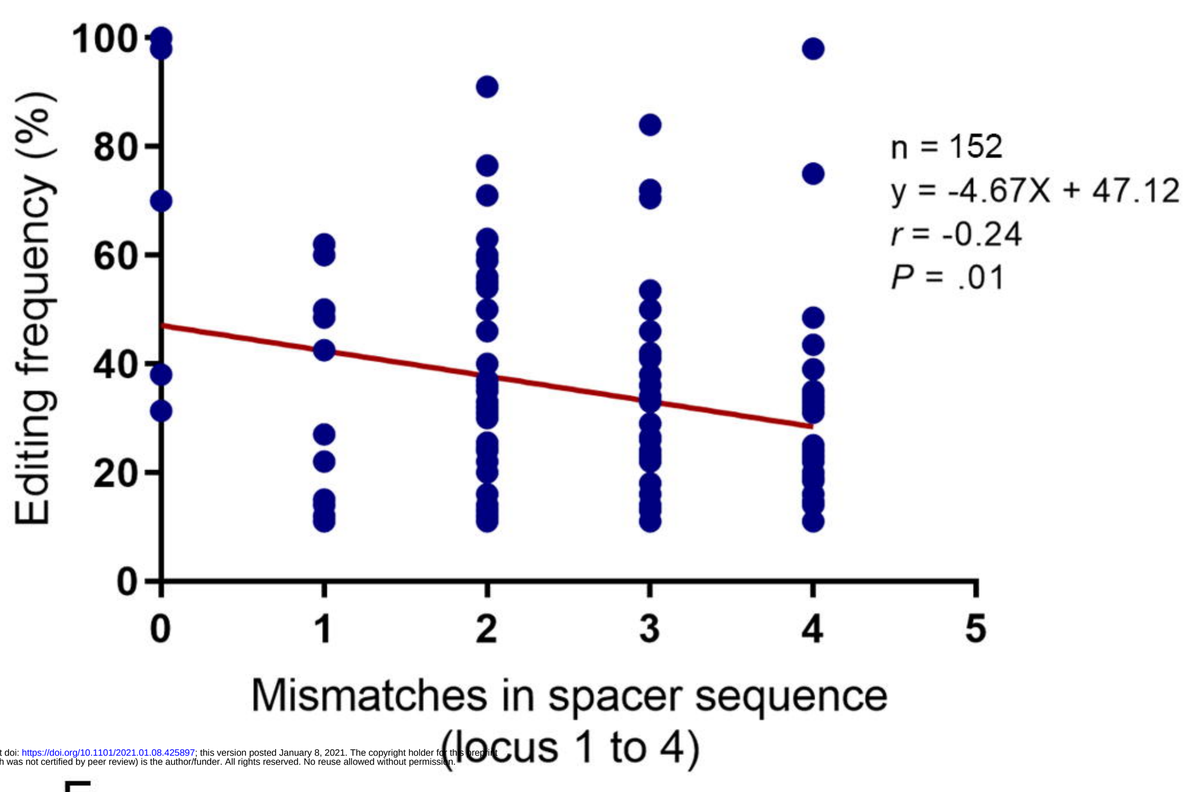


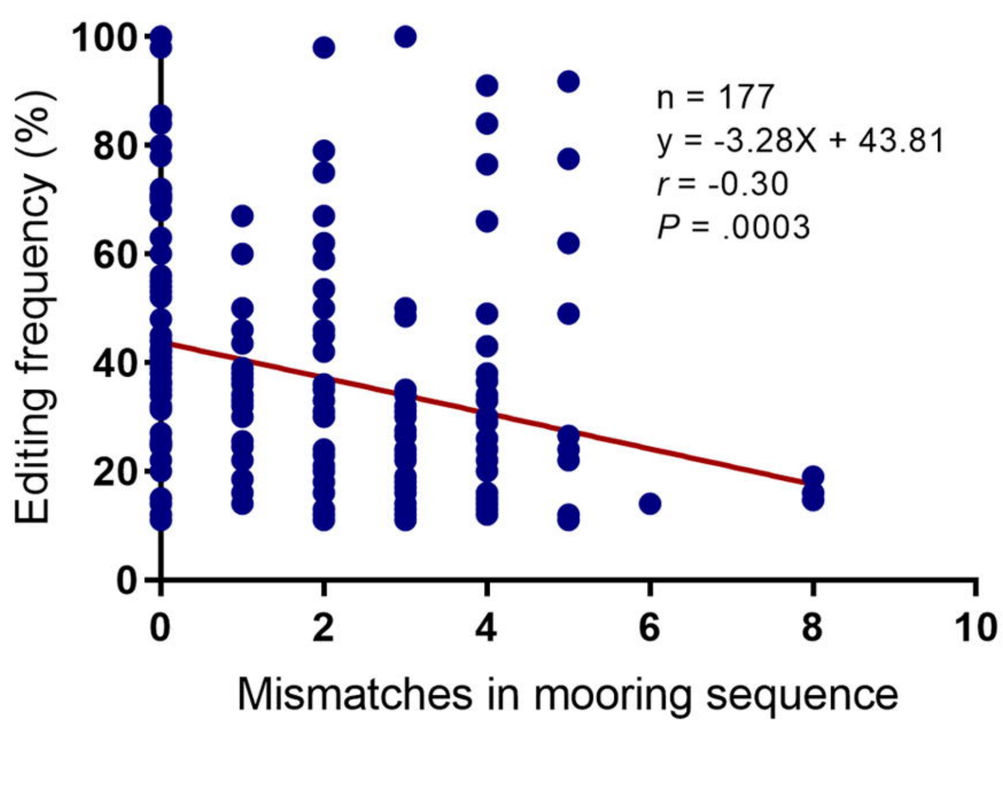
Figure 3











F

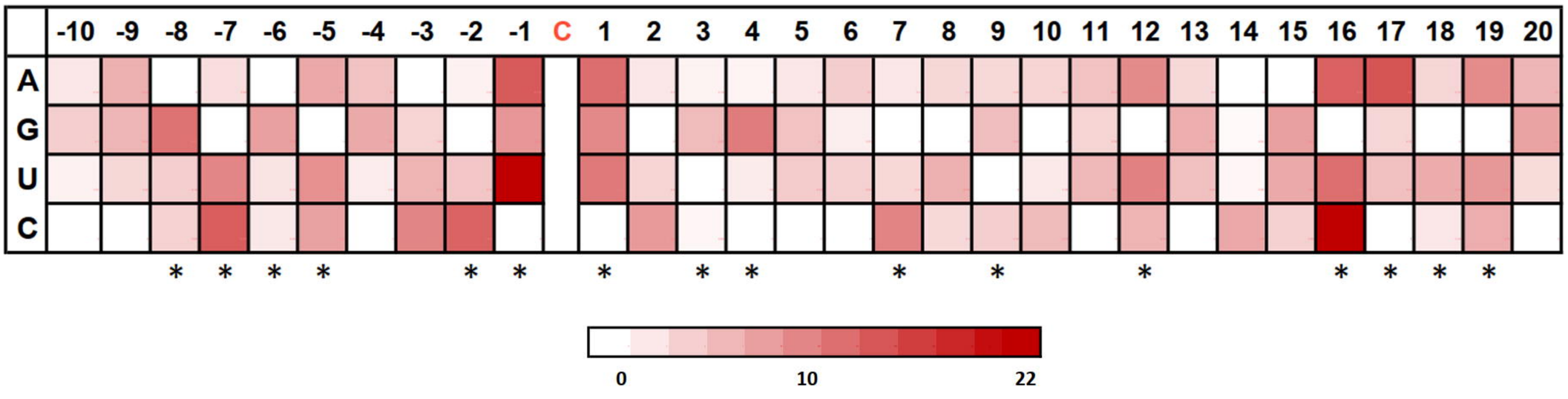
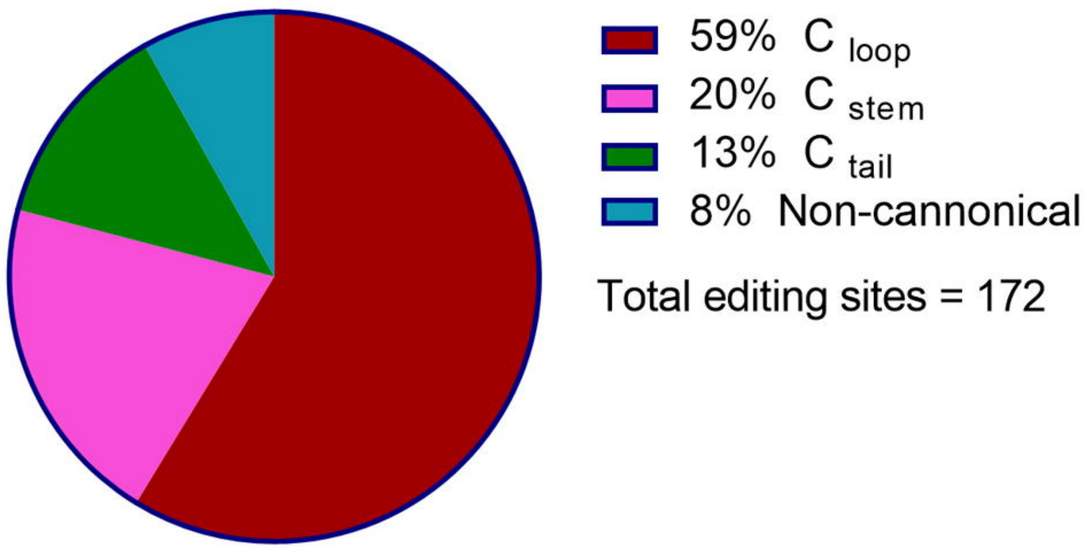


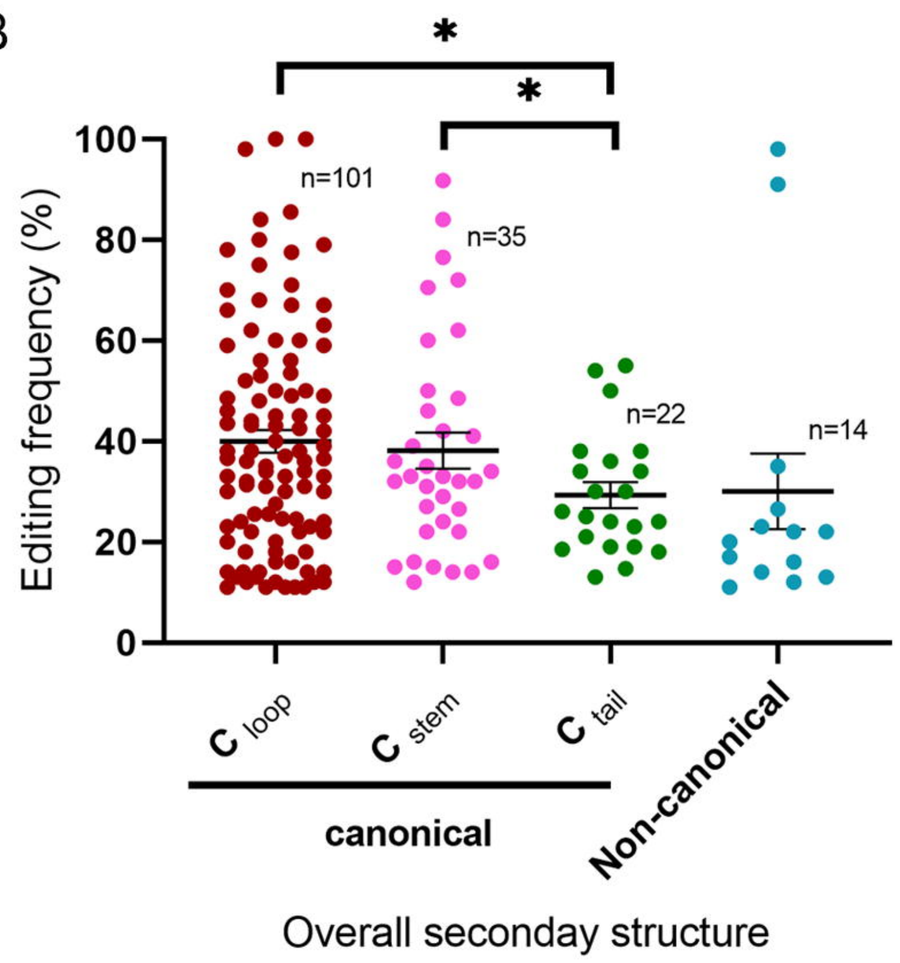
Figure 4

A



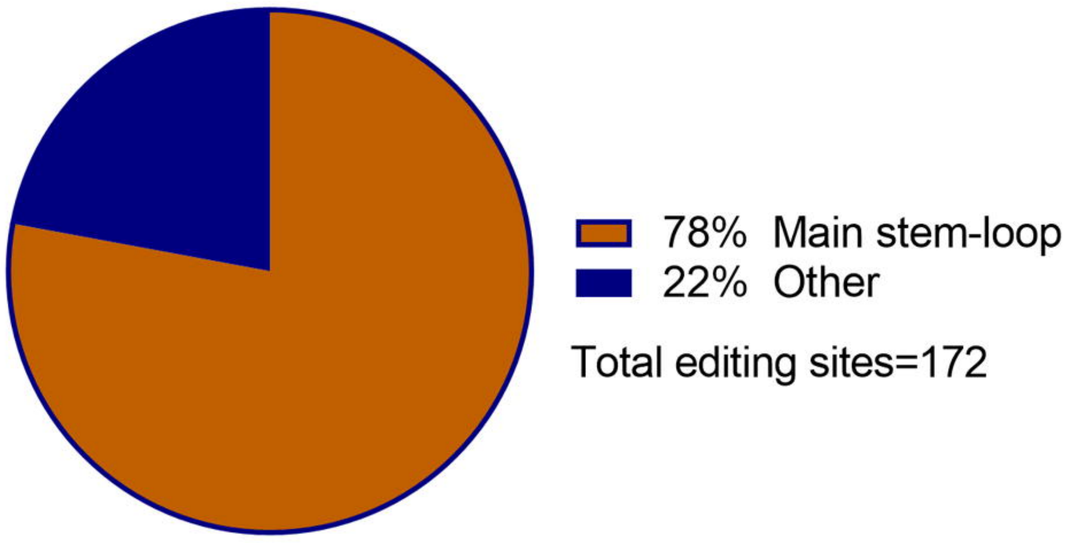
Overall secondary structure of C-to-U RNA editing sites

B



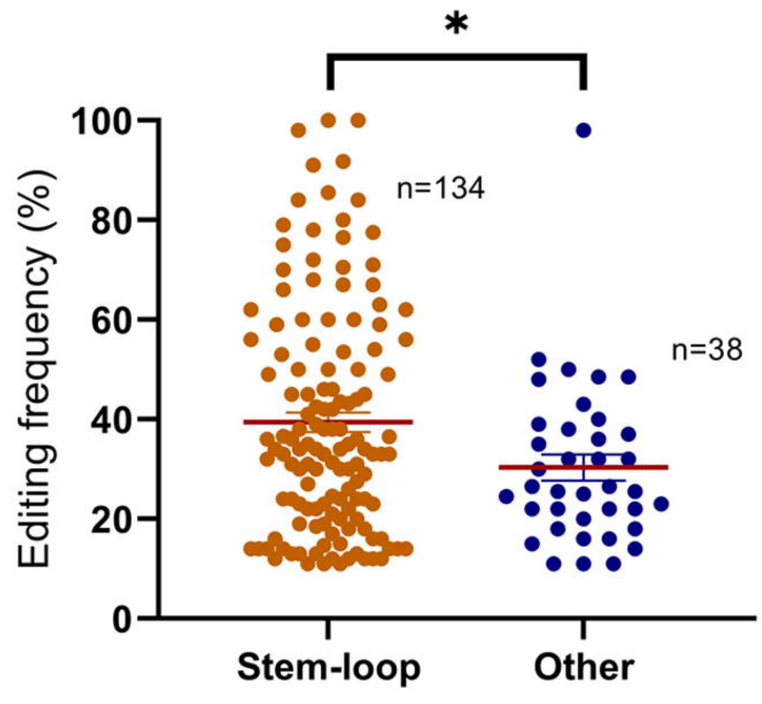
Overall seconday structure

C



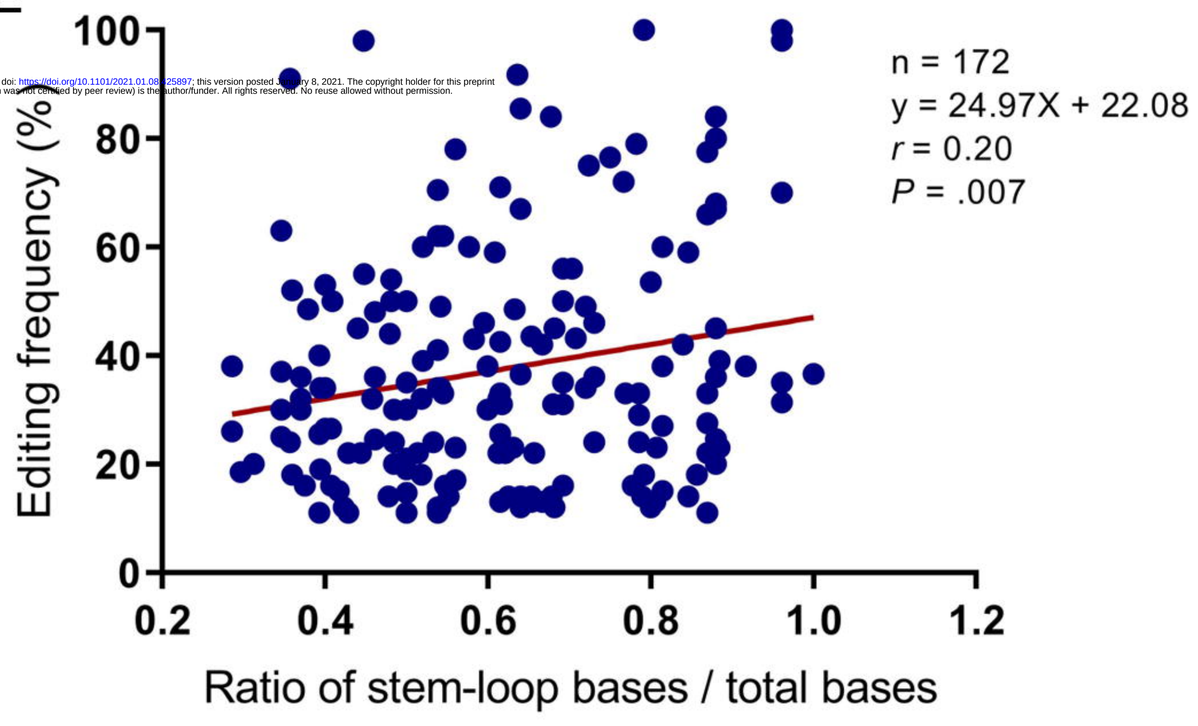
Location of mooring sequence in secondary structure of C-to-U RNA editing sites

D



Location of mooring sequence

E



F

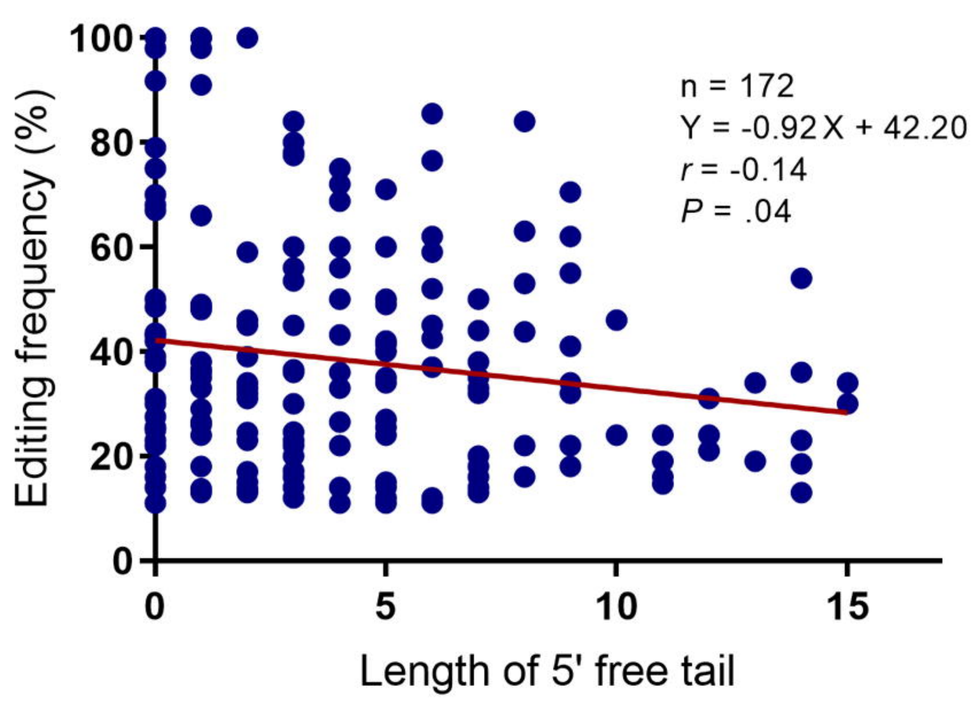
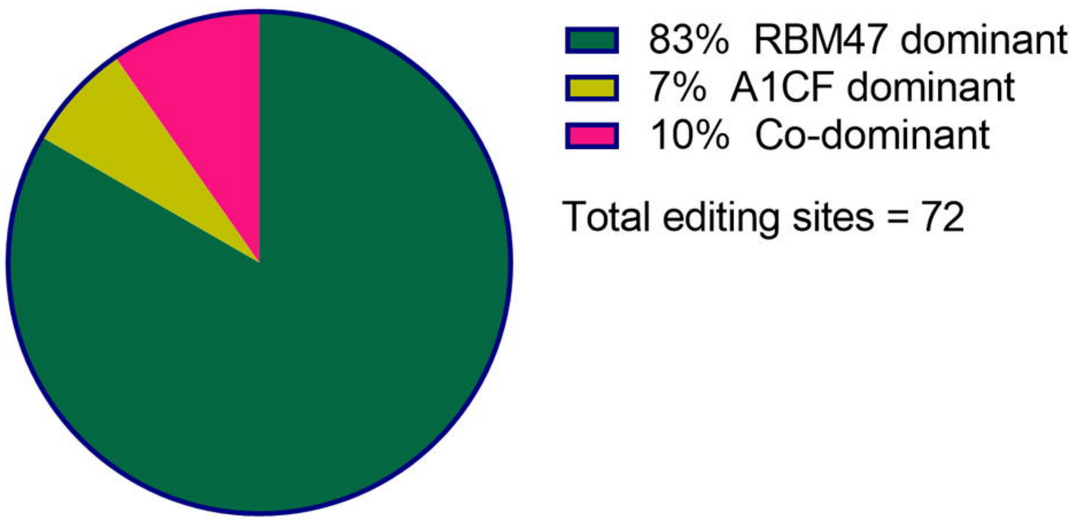


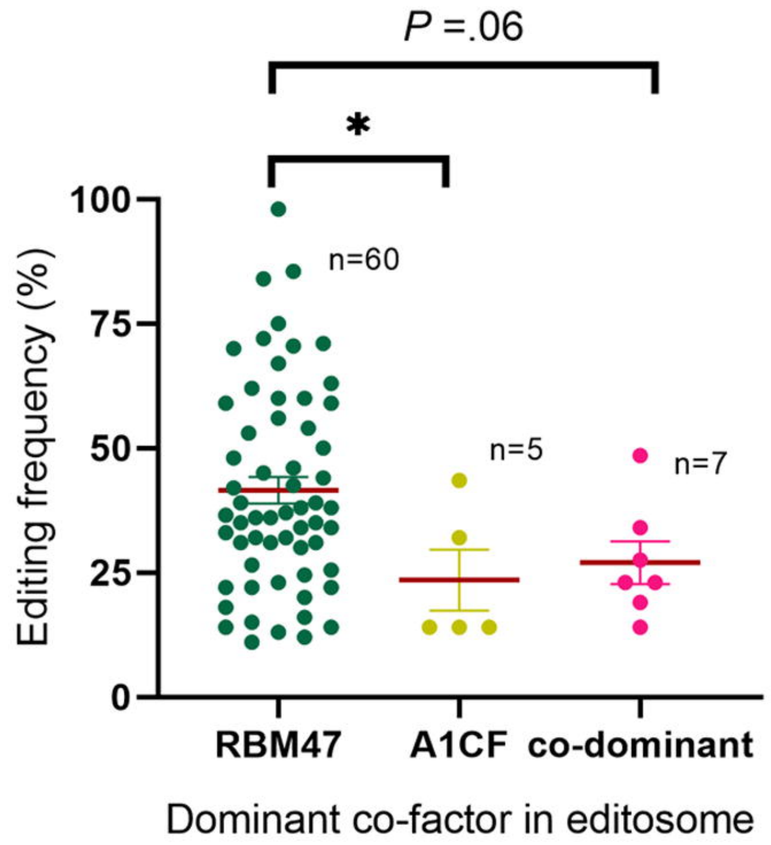
Figure 5

A



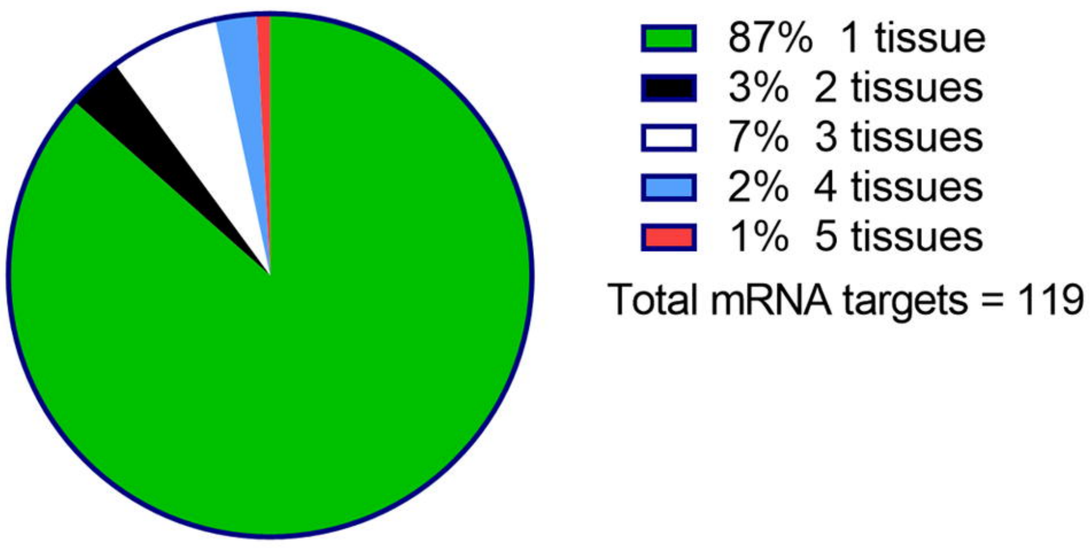
Co-factor group distribution of C-to-U RNA editing sites

B



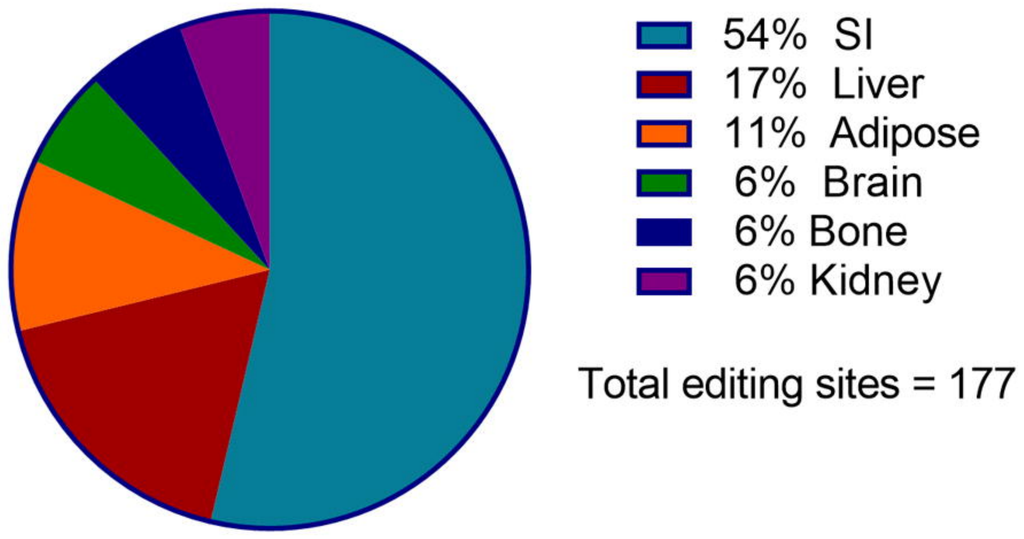
Dominant co-factor in editosome

C



Distribution of C-to-U RNA editing tissue number per target

D



Tissue distribution of C-to-U RNA editing sites

E

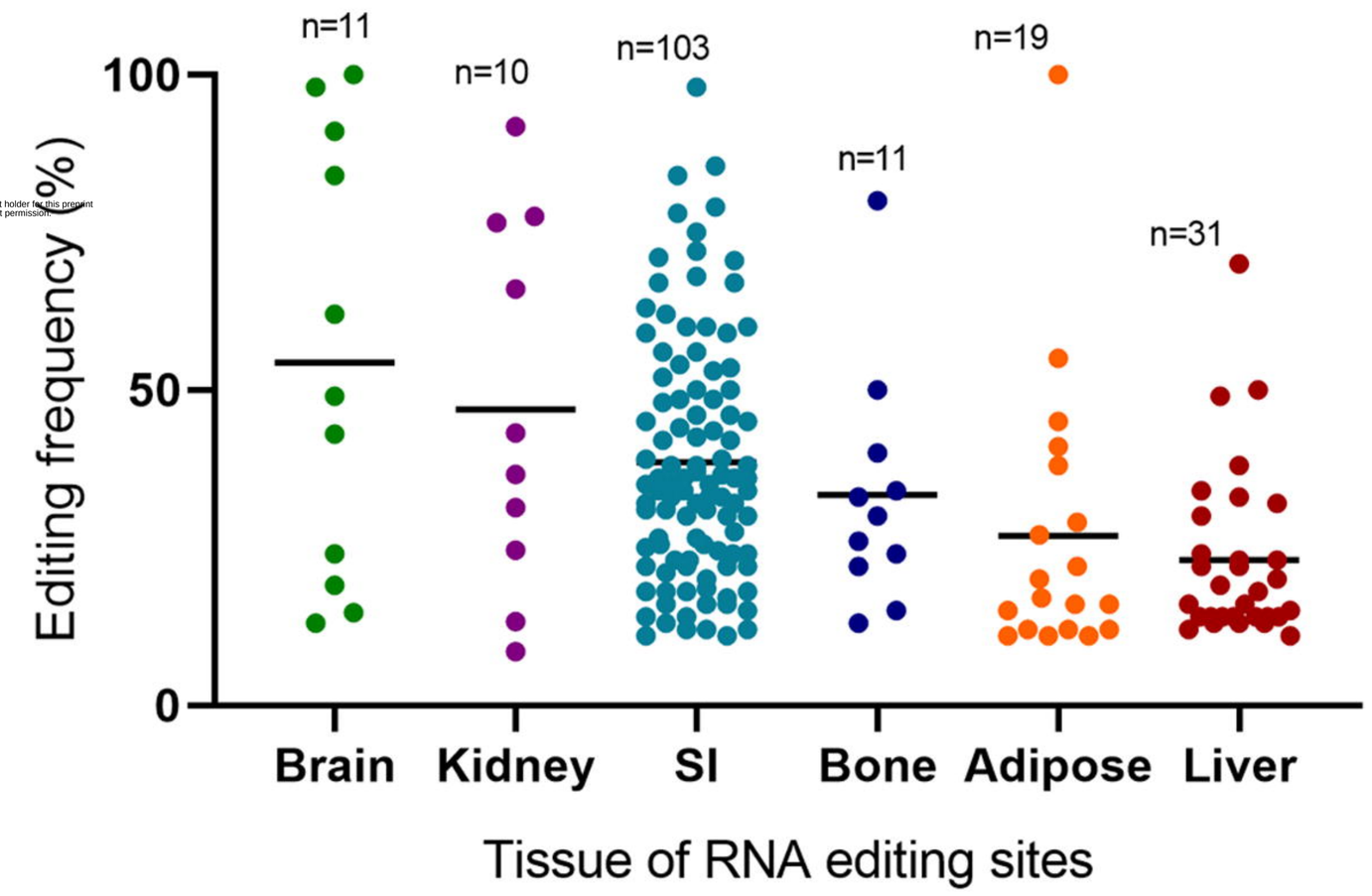
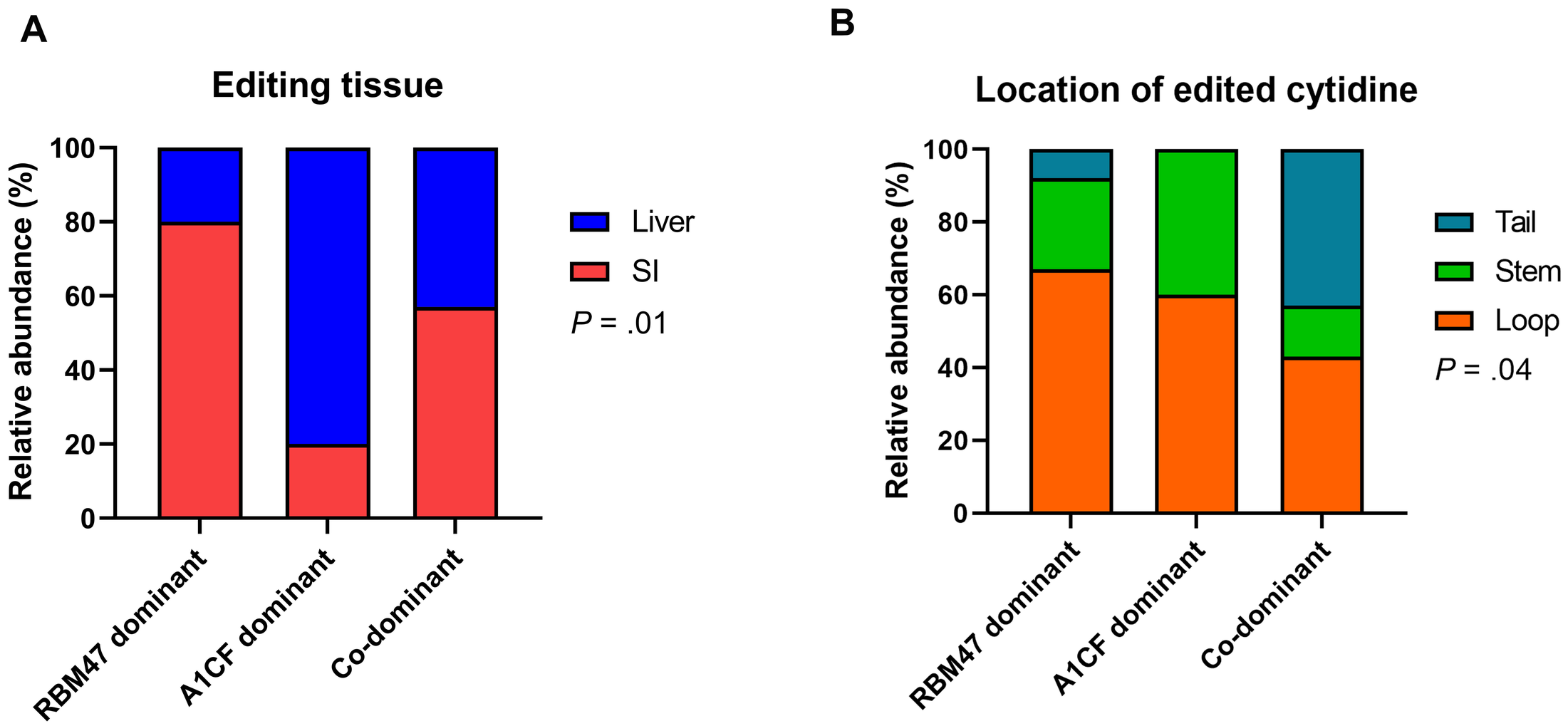
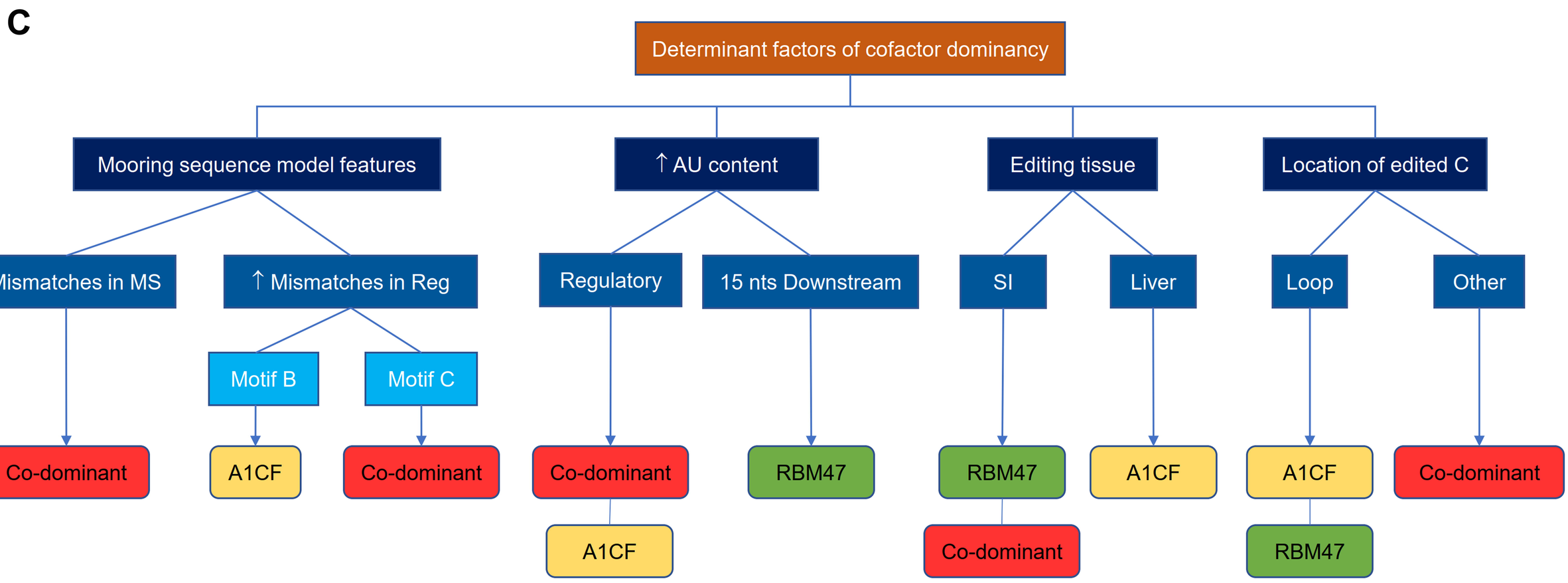


Figure 6


A
Editing tissue
Relative abundance (%)
100
80
60
40
20
0
Liver
SI
P = .01
RBM47 dominant
A1CF dominant
Co-dominant
B
Location of edited cytidine
Relative abundance (%)
100
80
60
40
20
0
Tail
Stem
Loop
P = .04
RBM47 dominant
A1CF dominant
Co-dominant


C
Determinant factors of cofactor dominancy
Mooring sequence model features
↑ AU content
Editing tissue
Location of edited C
↑ Mismatches in MS
↑ Mismatches in Reg
Regulatory
15 nts Downstream
SI
Liver
Loop
Other
Motif B
Motif C
Co-dominant
A1CF
Co-dominant
Co-dominant
RBM47
RBM47
A1CF
A1CF
Co-dominant
A1CF
Co-dominant
RBM47