Citation for published version:
Savisaar, R & Hurst, L 2017, 'Both maintenance and avoidance of RNA-binding protein interactions constrain coding region evolution', Molecular Biology and Evolution, vol. 34, no. 5, pp. 1110-1126.
https://doi.org/10.1093/molbev/msx061

DOI:
10.1093/molbev/msx061

Publication date:
2017

Document Version
Peer reviewed version

Link to publication

Article: Discoveries

Both maintenance and avoidance of RNA-binding protein interactions constrain coding sequence evolution

Savisaar, Rosina and Hurst, Laurence D.

Department of Biology and Biochemistry, The Milner Centre for Evolution, University of Bath, BA2 7AY, Bath, United Kingdom

Corresponding author: Rosina Savisaar (r.savisaar@bath.ac.uk)

Abstract

While the principal force directing coding sequence (CDS) evolution is selection on protein function, to ensure correct gene expression CDSs must also maintain interactions with RNA-binding proteins (RBPs). Understanding how our genes are shaped by these RNA-level pressures is necessary for diagnostics and for improving transgenes. However, the evolutionary impact of the need to maintain RBP interactions remains unresolved. Are coding sequences constrained by the need to specify RBP binding motifs? If so, what proportion of mutations are affected? Might sequence evolution also be constrained by the need not to specify motifs that might attract unwanted binding, for instance because it would interfere with exon definition? Here, we have scanned human CDSs for motifs that have been experimentally determined to be recognized by RBPs. We observe two sets of motifs – those that are enriched over nucleotide-controlled null and those that are depleted. Importantly, the depleted set is enriched for motifs recognized by non-CDS binding RBPs. Supporting the functional relevance of our observations, we find that motifs that are more enriched are also slower-evolving. The net effect of this selection to preserve is a reduction in the over-all rate of synonymous evolution of 2-3% in both primates and rodents. Stronger motif depletion, on the other hand, is associated with stronger selection against motif gain in evolution. The challenge faced by our CDSs is therefore not only one of attracting the right RBPs but also of avoiding the wrong ones, all while also evolving under selection pressures related to protein structure.

Introduction

One of the most captivating problems in molecular evolution is that of multiple coding − how the very same DNA sequence can contain several overlapping layers of information. This was once believed to primarily characterize viral genomes, where open reading frames (ORFs) routinely overlap (Barrell et al. 1976; Normark et al. 1983; Belshaw et al. 2007; Chirico et al. 2010). It is understood now, however, that not only are overlapping genes more common in vertebrates than previously believed (Veeramachaneni et al. 2004; Michel et al. 2012), other forms of multiple coding are near-ubiquitous (Itzkovitz et al. 2010; Lin et al. 2011; Shabalina et al. 2013; Pancsa and Tompa 2016). For example, protein-coding regions can overlap with transcription factor binding sites (Stergachis et al. 2013; Birnbaum et al. 2014) (although the functionality of the sites is contested (Xing and He 2015; Agoglia and Fraser 2016)), functional RNA secondary structures (Chamary and Hurst 2005; Meyer and Miklos 2005; Pedersen et al. 2006; Smith et al. 2013) and microRNA targets (Lewis et al. 2005; Hurst 2006; Forman et al. 2008; Fang and Rajewski 2011; Hausser et al. 2013; Liu et al. 2015). This means that the evolution of coding sequences (CDSs) is directed not only by selection pressures related to the structure of the protein encoded for but also by the need to preserve such overlapping regulatory information.

Here, we have examined one particular layer of information in CDSs, namely target sites to RNA-binding proteins (RBPs). A constantly changing assortment of RBPs accompanies the (pre-)mRNA transcript throughout its life and coordinates gene expression (Glisovic et al. 2008; Muller-McNicoll and Neugebauer 2013; Singh et al. 2015). Although many of these proteins interact preferentially with untranslated regions or introns (e.g. Licatalosi et al. 2008; Xue et al. 2009; Ince-Dunn et al. 2012), others primarily bind CDSs (e.g. Grellscheid et al. 2011; Änkö et al. 2012; Ascano et al. 2012). We have sought to quantify to what extent the evolution of CDSs is constrained by the need to preserve or to avoid interactions with RBPs.

To do so, we have studied the enrichment and conservation of particular *k*-mers within CDSs. At least some RBPs recognize and interact with particular (classes of) sequence motifs in the mRNA (Ray et al. 2013; Li et al. 2014). If such an RBP functionally binds within CDSs, then this should lead to the over-representation and excess conservation (compared to null/neutral expectations) of the relevant motifs. On the other hand, it is possible that target sites to other sequence-specific RBPs are avoided in CDSs if interactions between CDSs and those RBPs have deleterious consequences. For example, if an RBP that normally functions by binding introns bound a CDS, it could theoretically interfere with exon definition during pre-mRNA processing or simply constitute a waste of the protein. Such avoidance should manifest itself in the associated motifs being less frequent than expected by chance. The impact on evolutionary rates should be two-fold: on the one hand, the avoided motifs themselves are expected to be fast-evolving due to pressure to degrade them. On the other hand, those *k*-mers that are a short mutational distance away from an avoided motif should be under selection against substitutions that would give rise to the avoided motif.

Such patterns of enrichment and conservation have been studied extensively for exonic splice enhancers (ESEs). ESEs are short RNA motifs, enriched at exon ends, that promote splicing and are important for the correct identification of the splice sites in a wide range of multicellular organisms (Blencowe 2000; Fairbrother et al. 2004; Wu et al. 2005; Wang and Burge 2008; Warnecke et al. 2008; Cáceres and Hurst 2013). They are under purifying selection (Fairbrother et al. 2004; Carlini and Genut 2006; Parmley et al. 2006; Parmley et al. 2007; Ke et al. 2008; Sterne-Weiler et al. 2011; Cáceres and Hurst 2013), leading in human and mouse to an estimated reduction in the over-all rate of evolution at synonymous sites of about 1.9% − 4% (Parmley et al. 2006; Cáceres and Hurst 2013). There is evidence that the pressure to conserve ESEs may also have an impact on protein evolution: higher ESE density, as well as higher splice factor binding site density, have been found to correlate with increased protein disorder (Macossay-Castillo et al. 2014; Smithers et al. 2015). Moreover, Parmley et al. (2007) showed that amino acid composition at exon ends, where ESEs are most frequent, is biased towards residues that are encoded for by codons that are frequent in ESEs (for a case study, see Falanga et al. (2014)). More generally, there is evidence that the proportion of an mRNA that is within a short distance to a splice site (and therefore likely enriched in splice regulatory information) is one of the main determinants of how fast the corresponding protein evolves (Parmley et al. 2007).

Most ESE function can probably be explained by interactions with RBPs, notably SR proteins (Blencowe 2000; Zhou and Fu 2013). The work discussed above on the evolution of these motifs therefore constitutes a step towards understanding the evolutionary importance of RBP binding more generally. In the present study, we expanded the scope of the analysis from splice enhancement alone to all the functions CDS-RBP interactions may have (see section 1 of the *Discussion* for further consideration of the functions of RBPs).

We assembled a large set of *k*-mers that have been demonstrated experimentally to be recognized by various RBPs, and scanned human CDSs for hits. Note that we are concerned strictly with global biases on mRNA sequence evolution and not with predicting individual binding sites, a wholly separate problem that would require a different approach (see *Materials and Methods* for discussion). We found the motifs to be both more frequent and more conserved than would be expected by chance from their nucleotide composition. We estimate the net effect of the need to preserve them to be a decrease of *ca.* 2.4% in the over-all rate of evolution at human synonymous sites − an estimate that is in line with those produced previously for ESEs (Parmley et al. 2006; Cáceres and Hurst 2013). This might suggest that ESEs alone capture a large fraction of the selective pressures acting on motifs recognized by RBPs as a whole.

Importantly, the task facing CDSs appears to be not simply to maintain necessary RBP interactions but also to avoid inappropriate ones. Indeed, although the over-all effect is one of motif enrichment, there are also many RBPs whose putative target motifs are depleted compared to nucleotide-controlled null, and appear to be selectively avoided in CDS evolution. It is possible that these represent RBPs whose interactions with CDSs can have deleterious consequences, either because they actively interfere with gene expression or because they divert the protein away from functional binding sites in other transcript regions.

Results

1. Putative RBP target motifs are non-neutrally evolving in CDSs, leading to an over-all decrease of ~2.4% in the human rate of synonymous evolution.

1.1. Putative RBP target motifs are enriched over expected in CDSs.

Is the frequency of putative RBP target motifs in CDSs consistent with neutral expectations or are there deviations that would suggest the presence of selection? We retrieved data on the experimentally determined sequence specificities of human RBPs from several databases. This provided us with 114 RBPs, each one associated to a particular set of *k*-mers with *k* ranging from 5 to 12 (from now on these *k*-mers will be referred to as *RBP motifs*; Additional File 1). The motifs were pooled across all the sets, resulting in a final list of 1483 unique RBP motifs. The techniques used to determine these motifs vary widely, ranging from nuclear magnetic resonance based approaches (e.g. Garcia-Mayoral et al. 2008) to high throughput competition assays, such as RNAcompete (Ray et al. 2013). We next compiled a set of 10,337 full human intron-containing CDSs (concatenations of all the coding regions from the transcript variant with the longest CDS). To alleviate problems of statistical non-independence, the CDSs were clustered into families of paralogs (Additional File 2). In the analyses described below, statistics were either averaged within families or only a single randomly picked gene was considered from each, resulting in 5845 independent data points for each estimate (see *Materials and Methods* for further details).

We then scanned the CDSs for RBP motifs and calculated the motif density, that is to say, the fraction of the bases in a given CDS that overlapped with any of the motifs. The median density was ≈0.573 (Supplementary Spread Sheet 1 in Additional file 4), meaning that over half of the sequence in a typical human intron-containing CDS overlaps with one or more RBP motifs. Does this deviate from the density that would be expected by chance for a set of motifs of this size and of this base content? We generated 1000 sets of simulant motifs of the same size and roughly the same dinucleotide composition as the set of RBP motifs. We determined the density of the simulant sets in our sequences and observed that none of them had a median density as high as that observed with real RBP motifs. RBP motifs are therefore enriched in

CDSs with a p-value of ≈0.001 ($p = \frac{n+1}{m+1}$, where n is the number of simulant sets that present a median density as great as or greater than that observed with the real motif set and m is the total number of simulant sets). This is an indication that there could indeed be selection to preserve these motifs.

In order to quantify this enrichment, we can calculate a normalized density value for each gene ($ND = \frac{true\ density - mean\ of\ simulated\ densities}{mean\ of\ simulated\ densities}$). ND is a measure of enrichment over the nucleotide-controlled null. An ND value of 0 signifies that the motifs are about as frequent as would be expected by chance given their nucleotide composition, whereas an ND of 1 means that they are twice as frequent as expected and an ND of −0.5 that they are half as frequent. For RBP motifs, we recover a median ND value of ≈0.115.

1.2. RBP motifs are under purifying selection.

If the motif enrichment reported above truly reflects the functionality of (a subset of) the *k*-mers rather than, say, a methodological bias in the simulations, then in addition to being enriched, the motifs should also be slower-evolving than expected from their nucleotide composition. To test this prediction, we aligned the gene regions overlapping the motifs to the homologous regions in the macaque (*Macaca mulatta*) genome and calculated the rate of evolution at synonymous sites (d_S). We then applied the same procedure to each of the 1000 simulated versions of the RBP motifs set. This generated a distribution of simulant d_S estimates, from which we calculated an empirical conservation p-value ($p = \frac{n+1}{m+1}$, where n is the number of simulant sets that present a d_S as low as or lower than that observed with the real motif set and m is the total number of simulant sets) and a normalized d_S estimate ($normalized\ d_S = \frac{true\ d_S - mean\ of\ simulated\ d_S}{mean\ of\ simulated\ d_S}$). RBP motifs show a significant reduction in d_S (raw $d_S \approx 0.064$; normalized $d_S \approx -0.041$; $p \approx 0.003$). This suggests that CDSs are indeed under selection to preserve RBP motifs, underlining their functionality.

In order to further verify this result using a different method, we compared evolutionary rates at fourfold degenerate sites that overlapped RBP motifs to rates at those that did not, performing the analysis separately for each dinucleotide (see

Supplementary Text 1 in Additional File 5 for details). Although the effects recovered were weaker than those obtained in the d_S analysis reported above, the majority of dinucleotides do evolve more slowly within RBP motifs than elsewhere ($\chi^2 \approx 4$, $p < 0.05$ from χ^2 test; $p \approx 0.017$ from a paired one-tailed Wilcoxon signed rank test comparing the rates obtained for each dinucleotide in motifs and non-motifs). It appears therefore that our results cannot simply be due to a bias in the normalization procedure that would cause a few fast-evolving dinucleotides (such as *CG*) to be over-represented in simulants when compared to the true motifs.

1.3. RBP motif enrichment is stronger in genes that are expressed more tissue-specifically.

The hypothesis of RBP motif functionality possibly makes a further prediction, namely that the motifs should be enriched more in genes that are more highly expressed or expressed in a greater number of tissues. This is because various errors made during gene expression should have greater fitness consequences if the transcript is more abundant, assuming that all else is equal. Selection on regulatory signals that help ensure correct gene expression should therefore be stronger, leading to higher enrichment.

We obtained FANTOM5 expression data (Fantom Consortium et al. 2014) for the genes in our dataset. For each gene, we calculated the following expression parameters: expression breadth (fraction of tissues where the gene is expressed), median expression, maximum expression, and median expression in tissues where the gene is expressed (Supplementary Spread Sheet 18 in Additional File 4). After Bonferroni correction, we find that ND indeed correlates significantly with three of these variables (Table 1). However, contrary to our expectations, the sign of the correlation is negative rather than positive. In addition, the relevant parameter seems to be the number of tissues in which the gene is expressed more so than transcript abundance in any given tissue. In other words, it appears that the more tissue-specific a gene's expression pattern, the more RBP motifs are enriched. This might reflect greater levels of regulation in more narrowly expressed genes. This tendency must be stronger than any increased purifying selection on genes with greater expression breadth.

We were concerned that the negative correlation between ND and expression parameters could be reflecting properties of simulant motifs rather than of the true RBP motifs. Namely, the formula for calculating ND ($ND = \frac{true\ density - mean\ of\ simulated\ densities}{mean\ of\ simulated\ densities}$) requires one to divide by the mean of simulated densities. If simulated density correlated positively with expression breadth, this could lead to a negative correlation between ND and expression breadth without there being any relationship between true motif density and expression. We therefore repeated the analysis using *Z*-scores rather than ND ($Z = \frac{true\ density - mean\ of\ simulated\ densities}{standard\ deviation\ of\ simulated\ densities}$). *Z*-scores should be more robust to fluctuations in the simulated mean, as this parameter does not appear in the denominator. It is therefore reassuring that we observed a negative correlation between *Z* and expression breadth ($\rho \approx -0.156$, $p < 2.2 * 10^{-16}$; Spearman rank correlation). In addition, raw motif density also correlates negatively and significantly with expression breadth ($\rho \approx -0.123$, $p < 2.2 * 10^{-16}$; Spearman rank correlation), demonstrating that the effect we observe for ND cannot be explained solely by patterns of simulated density.

To conclude, although the sign of the correlation is different from what was hypothesized, the fact that RBP motif enrichment correlates significantly with expression parameters adds further support to the claim that these motifs are functional in CDS.

1.4. The need to preserve RBP motifs leads to an over-all reduction of ~2–3% in primate and rodent $\mathbf{d_s}$.

It has been estimated (Parmley et al. 2006; Cáceres and Hurst 2013) that the need to preserve ESEs causes a reduction of about 1.9% – 4% in the over-all rate of evolution at synonymous sites (d_S). What would be the analogous estimate for RBP motifs? To find out, one can multiply normalized d_S by ≈0.573, that is to say, the fraction of the sequence in the median human CDS that is made up of RBP motifs. This provides us with an estimate for the over-all reduction in the d_S of the median gene that can be attributed to the pressure to preserve RBP motifs. This statistic turns out to be ≈−0.024. It therefore appears that the need to preserve RBP motifs indeed places a weak but detectable constraint on sequence evolution within human protein-coding regions. The magnitude of the effect we report for RBP motifs in CDSs is in line with previous estimates obtained for ESEs. However, not all RBP motif related constraint seems to be splice-associated: the net decrease in d_S is similar between intron-containing and intronless sequences (Supplementary Text 2 in Additional File 5, Supplementary Spread Sheet 7 in Additional File 4, and Additional File 3), suggesting that splicing-independent factors are important in directing RBP motif evolution.

We next asked whether our results concerning selection on RBP motifs in CDSs could be confirmed in another system. We repeated the analysis on 15,631 mouse (*Mus musculus*) CDSs, using motifs derived for mouse RBPs (Additional File 1; Supplementary Spread Sheet 10 in Additional File 4). We employed the rat (*Rattus norvegicus*) genome for estimating conservation. We recovered a lower median motif density than in human (≈0.339 and ≈0.573, respectively). However, this is likely simply because the set of motifs was smaller in mouse (736 motifs compared to the 1483 in human). The extent of enrichment (ND ≈ 0.128; $p \approx 0.010$) was similar to that obtained in human. Excess conservation was slightly more pronounced (raw $d_S \approx 0.165$; normalized $d_S \approx -0.063$; $p \approx 0.010$), leading to an estimate of ≈2.1% for the over-all reduction in d_S that would be due to the need to preserve RBP motifs. Data from mouse therefore also provides evidence for purifying selection on RBP motifs, and leads to similar conclusions with regards to the magnitude of this constraint.

1.5 RBP motif related constraint is as strong in CDSs as it is in introns and untranslated regions (UTRs).

We have provided evidence that RBP motifs are under selection in CDSs. However, is the over-all evolutionary impact of this selection substantially weaker in CDSs than in the non-coding regions of protein-coding genes? This might be expected as the latter regions are not under the additional constraint of specifying protein structure. They could therefore be particularly prone to the accumulation of regulatory signals, such as RBP binding sites. We analysed RBP motif density and conservation in 5’UTRs, 3’UTRs, full introns and exon proximal intronic regions (the 100 bp immediately upstream or downstream from an exon; Supplementary Spread Sheets 13-17 in Additional File 4). We found evidence for RBP motif conservation in all compartments and in all bar the intronic sequence from the downstream flanks of exons the effect was significant (Table 2).

Contrary to our expectations, the over-all constraint (the product of the motif density and the nucleotide-normalized conservation estimate) was stronger in CDSs than in any of the non-coding regions (Table 2). This could be reflecting the fact that synonymous sites are not subject to selective pressures related to amino acid sequence. The selection acting on non-coding signals in CDSs could therefore be disproportionately concentrated at synonymous sites, leading to a strong effect at the level of d_S. However, any such reasoning should be taken with a grain of salt as the conservation statistics were obtained slightly differently for CDSs and for the other sequence regions (using PAML *codeml* for CDSs and PAML *baseml* for the non-coding sequence). We therefore merely note that we find no evidence for unusually weak RBP motif related constraint in CDSs, and refrain from drawing conclusion from more fine-scale comparisons.

In conclusion, we have attempted to quantify the extent to which excess conservation at RBP motifs leads to a global decrease in d_S. We have found this figure to be about 2.4% − approximately the same level of constraint as can be observed in the non-coding regions of protein-coding genes. We emphasize that the figures we provide are to be taken as rough estimates only, as they are sensitive to the number of motifs defined as RBP motifs and to the procedure used for calculating the neutral expectation. Note also that our approach does not discriminate between strong selection acting on a few of the motifs in our set and weak selection acting on many. In the sections to follow, we will attempt to clarify this issue.

2. Nucleotide-controlled density varies greatly among motifs putatively recognized by different RBPs, with depletion no less frequent than enrichment.

2.1. When RBP motifs are grouped based on the cognate RBP, the enrichment **p**-*values of the resulting motif sets distribute bimodally.*

We determined above that RBP motifs were both more frequent and more conserved in CDSs than expected from their nucleotide composition, leading to a slight decrease in over-all d_S. It remains unclear, however, what the contributions of the motifs putatively recognized by different RBPs are to this result. Are more or less all RBP motifs enriched over expected or is the over-all enrichment largely driven by a subset of the motifs? Could some RBP motifs be not enriched but depleted instead? For instance, an intronic splice regulator binding within an exon could hypothetically interfere with exon recognition and so the presence of the cognate motifs within exons might be deleterious.

We repeated the analysis of motif density but instead of pooling the motifs, we considered the *k*-mers associated to each RBP separately. From here on, we will use the phrase *motif set* to refer to the motifs putatively recognized by a particular RBP. In total, there are therefore 114 motif sets, each corresponding to one RBP (see Supplementary Spread Sheet 19 in Additional File 4 for the sizes of the motif sets). As above for the pooled analysis, we generated 1000 approximately dinucleotide-matched simulated versions of each motif set so that we could calculate ND and an enrichment *p* for the motifs putatively recognized by each RBP (Supplementary Spread Sheet 2 in Additional File 4).

Some motif sets were very rare, leading to concerns over the reliability of estimating ND and other parameters in such cases. Because of this issue, we removed motif sets where hits to neither the true motifs nor the simulant sets reached a pre-defined density threshold (see *Materials and Methods*). After this filtering step, 81 motif sets remained (Additional File 1), containing a total of 1213 unique motifs. The enrichment *p*-values obtained for most of them were non-significant. However, there was a peak at either extreme (near 0 and near 1) when they were plotted out as a histogram (Figure 1A), leading to a significantly non-unimodal distribution ($D \approx 0.069$; $p \approx 0.005$; Hartigans' dip test).

In other words, a large proportion of the motifs fall into one of two classes: a (near-)significantly enriched class and a (near-)significantly depleted class. The over-all enrichment over expected that is obtained when all the motifs are pooled is therefore the average of many competing trends: the motifs putatively recognized by some RBPs are enriched, whereas others distribute at random frequencies or are altogether depleted.

2.2. The bimodal distribution of enrichment **p**-*values is specific to RBP motifs.*

Is this bimodal distribution of *p*-values specific to RBP motifs or could it be an artefact of our method for estimating *k*-mer enrichment? In the latter case, a similar distribution of *p*-values should also occur with motifs that are not expected to be biologically meaningful. We therefore replaced each motif within each motif set with a random *k*-mer of the same length and repeated the density analysis with these random motifs. We then generated 1000 sets of approximately dinucleotide-matched simulant motifs for each random motif set in order to calculate the enrichment *p*-values, identically to the analysis performed above for RBP motifs.

Unlike the RBP motifs, the random motifs showed no tendency for extreme *p*-values (black line in A; Supplementary Spread Sheet 3 in Additional File 4). To formally confirm this visual observation, we classed the *p*-values into two groups: below 0.1 or above 0.9, and between 0.1 and 0.9 (included). We then counted the number of *p*-values in either group and found the proportion to be significantly different for the RBP motifs and for random *k*-mers (χ^2 = 75.593, p < 0.001). In order to test the significance of the depletion effect specifically, we also compared the proportion of *p*-values above 0.9 to those below or equal to 0.9 for RBP motifs and for random *k*-mers. This difference was also significant (χ^2 = 132.819, p < 0.001). The bimodal distribution of enrichment *p*-values is therefore unlikely to result from methodological biases. We also considered the possibility that differences in stop codon content between the motifs and their simulants could be contributing to the depletion observed. The details of this analysis can be found in Supplementary Text 3 in Additional File 5, and Supplementary Figures 2 and 3, also in Additional File 5. Briefly, we found that although this factor might play some role in determining ND, it does not seem to explain the over-all pattern.

In conclusion, the tendency for extreme enrichment *p*-values exhibited by RBP motif sets is probably not due to methodological factors, as control motifs not thought to be biologically meaningful do not display this pattern. It is therefore likely that it is a reflection of the functionality of at least some of the motif sets.

3. The variation in enrichment between different sets of RBP motifs likely reflects functional differences.

3.1. Motif sets that are more strongly enriched also tend to be more conserved.

We have seen above that the extent of enrichment varies between sets of motifs putatively recognized by distinct RBPs. If this variation reflects differences in the functional importance of the motifs, then it should correlate with evolutionary rate: those motif sets that are more enriched should also be more conserved. To test this prediction, we calculated d_S, normalized d_S and a conservation *p*-value separately for each motif set (Supplementary Spread Sheet 4 in Additional File 4). As predicted under a functional hypothesis, we recovered a significant correlation between a motif set's ND and its normalized d_S ($\rho \approx -0.507$; $p \approx 1.388 * 10^{-6}$; Spearman rank correlation; Figure 2A; see Supplementary Figure 4 in Additional File 5 for qualitatively similar results obtained using enrichment *Z*-scores instead of ND, which controls for differences in the variance of the simulated density values; see Supplementary Figure 5 in Additional File 5 for a version of Figure 2 where each data point is labelled according to the associated RBP). Similarly, there is a significant positive correlation between enrichment *p*-values and conservation *p*-values ($\rho \approx 0.503$, $p \approx 1.75 * 10^{-6}$; Spearman rank correlation). The variation in the extent of enrichment, therefore, indeed likely results from functional differences between sets.

We repeated this analysis also for intronless CDSs and recovered similar patterns to those observed in intron-containing ones, once again underscoring the importance of processes other than splicing for determining RBP motif usage and evolution (Supplementary Text 2 in Additional File 5, and Supplementary Spread Sheets 8 and 9 in Additional File 4). We also performed the analysis using mouse CDSs and mouse RBPs (Supplementary Spread Sheets 11 and 12 in Additional File 4; Additional File 2). Like in human, we obtained a significant negative correlation between ND and normalized d_S ($\rho \approx -0.312$; $p \approx 0.005$; Spearman rank correlation), and a significant positive correlation between enrichment *p*-values and conservation *p*-values ($\rho \approx 0.352$; $p \approx 0.001$; Spearman rank correlation).

It could be pointed out that there is a significant correlation between the ND and the raw density of motif sets ($\rho \approx 0.292$, $p \approx 0.008$; Spearman rank correlation), and that the reliability of estimated d_S values is expected to depend on the amount of information available, which in its turn depends on the raw density. Therefore, the correlation between ND and normalized d_S could be due to less noisy estimation of normalized d_S in motif sets with greater ND. This is worrying because raw density is partially determined by methodological factors, such as the number of motifs in the set ($\rho \approx 0.674$, $p \approx 5.515 * 10^{-12}$; Spearman rank correlation between motif number and raw density) and the length of the motifs ($\rho \approx -0.323$, $p \approx 0.003$; Spearman rank correlation between median motif length and raw density). However, this alternative explanation predicts that in addition to the negative correlation between ND and normalized d_S, there should also be one between raw density and normalized d_S. This prediction is incorrect: there is no significant correlation between the raw density of a motif set and its normalized d_S ($\rho \approx 0.007$, $p \approx 0.949$; Spearman rank correlation). This confound is therefore unlikely to explain our results. We also note that several of the motif sets that present particularly extreme values for both ND and for normalized d_S are composed of very few motifs (see, for instance, CUGBP, Elav-Like Family Member 1 (CELF1) and Sterile Alpha Motif Domain Containing 4A (SAMD4A) in Supplementary Figure 5 in Additional File 5) and might therefore give rise to less reliable estimation of normalized d_S. Could our results be due to the presence of noisy outliers? This does not seem to be the case: we repeated the analysis after having removed all motif sets with fewer than 5 motifs and the significant correlation between ND and normalized d_S remained ($\rho \approx -0.520$, $p \approx 5.773 * 10^{-4}$; Spearman rank correlation).

It therefore appears likely that the motif sets that show the strongest enrichment are those recognized by RBPs whose interactions with CDSs are the most important to maintain. Do the associated RBPs also show preferential binding in CDSs in experimental studies? We annotated the RBPs as either *CDS-binding*, *non-CDS-binding* or *unknown* based on published high-throughput crosslinking and immunoprecipitation studies (CLIP-Seq) (Licatalosi et al. 2008; Xue et al. 2009; Hafner et al. 2010; Konig et al. 2011; Van Nostrand et al. 2016) (see Supplementary Spread Sheet 5 in Additional File 4 for references to data sources). The motif sets that were associated with CDS-binding RBPs indeed had greater raw density ($p \approx 0.016$; one-tailed Mann-Whitney *U*-test), greater ND ($p \approx 0.006$; one-tailed Mann-Whitney *U*-test) and lower enrichment *p*-values ($p \approx 0.009$; one-tailed Mann-Whitney *U*-test;

Figure 1B) than those annotated as non-CDS-binding. This concordance with experimental data both lends credence to the motif to RBP mapping and provides further support for the functional relevance of our observations.

In summary, the motif sets that are more strongly enriched in CDSs also tend to be slower-evolving, suggesting that they represent a subset of RBP motifs whose presence in CDSs has particular functional importance.

3.2. The depletion of certain motif sets is likely due to purifying selection to avoid them.

We noted above that despite the over-all enrichment of RBP motifs over nucleotide-controlled null in CDSs, many of the motif sets associated to individual RBPs were depleted instead. As this depletion is not observed for random *k*-mers (black line in Figure 1A), it most likely reflects selection to avoid motifs recognized by RBPs whose interactions with CDSs can be deleterious, either because they constitute a waste of the protein on inappropriate binding or because they interfere with gene expression. The latter type of scenario is easy to imagine in the case of splicing: an exon is partially defined by the factors that bind to it and so a change in the complement of binding partners could hypothetically interfere with exon recognition.

The implications of this avoidance for CDS evolution are likely two-fold. Firstly, one expects purifying selection against the avoided motifs, resulting in a general constraint on the sequence space available in CDS evolution. A read-out of this effect would be a rarity of substitutions that give rise to an avoided motif. Secondly, when the avoided motifs do occur, there should be positive selection to degrade them. They should therefore be faster-evolving than expected from their nucleotide composition. The magnitude of the second selection pressure will depend on the efficiency of the first: if the purifying selection against the avoided motifs is sufficiently strong, then they might almost never go to fixation in a context where their presence is deleterious. For instance, it could be that the majority of the hits observed for such motifs are in locations where the local mRNA secondary structure prevents the RBP from accessing the site and so these motifs, although present in the sequence, would very infrequently actually interact with the RBP. In this case, no positive selection to lose the motifs is expected and the avoided motifs should instead be neutrally-evolving. It is also possible that certain RBP-CDS interactions, although deleterious in most cases, can be adaptive when they occur at very specific locations. In this latter scenario, the avoided motifs would be rare but under purifying selection when present.

In order to determine whether there was any evidence for selection to degrade certain motifs, we pooled the motifs from those sets whose enrichment *p*-value was above 0.9 in intron-containing CDSs (the strongest candidates for being avoided; from here on, we will refer to these motifs as the *depleted group*) and calculated their density and rate of evolution in intron-containing CDSs. This resulted in a set of 432 motifs with a median density of ≈0.069, a median ND of ≈−0.130 and an enrichment *p*-value of 1 (i.e. significant depletion). There is no evidence for positive selection on the motifs: rather, they are evolving at roughly the rate that would be expected by chance from their nucleotide composition (raw $d_S \approx 0.068$; normalized $d_S \approx 0.011$; conservation $p \approx 0.600$). This might suggest that purifying selection to avoid the motifs is sufficiently efficient to mostly prevent them from going to fixation at locations where their presence is deleterious. It is also possible, however, that because of the rarity of the depleted group motifs, we simply lack the power to pick up on any positive selection that is occurring. Moreover, if the avoidance only concerns certain gene regions, this might dilute any signal of positive selection further. For instance, the avoidance might be stronger in the outer regions of exons (the exon flanks) than at the very centre, as the flanks appear to be more crucial for splice regulation. This is evidenced by their enrichment in both splice-altering (Woolfe et al. 2010) and pathogenic (Wu and Hurst 2016) single-nucleotide polymorphisms.

To test this hypothesis, we extracted 69 base pairs from the extreme 5' end, the extreme 3' end and the very centre of 4563 human internal coding exons and calculated the d_S of the depleted group. In the 5' flank, depleted group motifs are indeed evolving faster than expected from their base composition but this effect is non-significant (5' flanks: raw $d_S \approx 0.073$; normalized $d_S \approx 0.116$; conservation $p \approx 0.915$). In exon cores and 3' flanks, however, the same motifs are evolving at chance rates (cores: raw $d_S \approx 0.068$; normalized $d_S \approx 0.024$; conservation $p \approx 0.635$; 3' flanks: raw $d_S \approx 0.072$; normalized $d_S \approx 0.045$; conservation $p \approx 0.727$). Given the non-significance of the effects, it appears that even when considering the different exonic sub-regions separately, there is little evidence for increased rates of evolution in regions overlapping depleted group motifs.

We next sought to directly test for purifying selection against the depleted group. We determined all four-fold degenerate sites in our set of intron-containing CDSs such that a single base substitution at the site would give rise to one of these motifs. We then aligned the CDSs to macaque orthologs and found that at ≈1.4% of such sites, the base that would create a depleted group motif were it used at that position in human was indeed present at the orthologous site in the macaque sequence (the site counts have been weighted based on site degeneracy − see *Materials and Methods*). We repeated the same analysis on 1000 sets of dinucleotide-matched simulant motifs and found the corresponding percentage to be ≈1.6% on average. This difference is slight but significant (one-tailed empirical $p \approx 0.009$ from the distribution of simulant values). This is evidence for selection against substitutions that would give rise to a depleted group motif.

Another way to test for purifying selection against certain RBP motifs is to consider the variation among motif sets. If motif depletion is largely driven by purifying selection to avoid, it is expected that the more a motif set is depleted, the more motif gain is selected against over evolution. To test this hypothesis, we repeated the analysis of sites that are a single substitution removed from a motif but this time separately for the individual RBP motif sets (Supplementary Spread Sheet 6 in Additional File 4). For each motif set, we calculated the fraction of one-removed sites where the base that would give rise to one of the motifs in the set in human was present in macaque. We then normalized this statistic by subtracting from this value the mean fraction observed for simulated sets and dividing the difference by the simulated mean. We then calculated the correlation between these normalized fractions and ND. As predicted, this correlation was significantly positive ($\rho \approx 0.538$, $p \approx 3.530 * 10^{-7}$; Spearman rank correlation; Figure 2B). Analysis of individual motif sets therefore also provides evidence that the depletion of certain motif sets is due to purifying selection to avoid them.

The fact that CDSs co-exist in the cell with RBPs therefore has the effect of carving out a sub-region of sequence space within which CDSs preferentially evolve.

Deviating from these constraints may not only lead to the loss of necessary CDS-RBP interactions but might also provoke inappropriate ones.

Discussion

1. An estimate for the decrease in the synonymous rate of evolution that is due to selection to preserve interactions with RNA-binding proteins.

In this study, we have sought not simply to test whether the need to preserve RBP binding constrains CDS evolution but also to quantify the evolutionary impact of any such dual coding. We estimate that the need to conserve motifs putatively recognized by RBPs leads to a decrease of *ca.* 2–3% in the over-all rate of evolution at synonymous sites in both primates and rodents compared with a nucleotide controlled null. This reduction in evolutionary rate, however, is not distributed uniformly across the RBP motifs, appearing to be driven by a subset of the motifs that are particularly enriched and conserved, while others occur at chance frequencies or are altogether depleted. Note also that the very low figure that we provide for the over-all decrease in evolutionary rates is likely an underestimate because the nucleotide-controlled null has been intentionally designed to be conservative. It is possible that some of the control sites overlap with functional RBP targets and are therefore conserved, leading to an overly low expected rate of evolution.

The estimate that we have produced for RBP motifs is comparable to the 1.9%−4% range that can be deduced from similar analyses on exonic splice enhancers (ESEs) (Parmley et al. 2006; Cáceres and Hurst 2013). This might indicate that ESEs alone capture a large fraction of the selective pressures acting on putative RBP target motifs more generally. This should not be taken to imply that all RBP-related constraint is due to the need to ensure correct splicing: we found both the over-all level of constraint, as well as the enrichment and conservation patterns of the individual sets of motifs putatively recognized by particular RBPs, to be remarkably similar between intron-containing and intronless sequences (Supplementary Text 2 in Additional File 5). This suggests that splicing-independent factors may be surprisingly important in shaping the RBP motif content of CDSs. This result concords with previous findings that ESEs are both enriched and conserved (compared to nucleotide-controlled null) also in genes that do not undergo splicing, indicating that they too might be relevant to processes other than splicing (Pozzoli et al. 2004; Savisaar and Hurst 2016).

Our data do not inform us on which particular splicing-independent functions might be the most relevant in directing RBP motif evolution. However, it is well established that RBPs that bind the CDS can indeed have such roles. For instance, the serine-arginine rich splice factor 1 (SRSF1) is crucial for maintaining genome stability (Li and Manley 2005; Tuduri et al. 2009), whilst the serine-arginine rich splice factors 3 and 7 (SRSF3 and SRSF7) have been shown to act as adapters in the nucleo-cytoplasmic transport of the intronless *H2a* mRNA (Huang and Steitz 2001; Huang et al. 2003). Other RBPs, such as fragile X mental retardation 1 (FMR1) (Kao et al. 2010) and Heterogeneous Nuclear Ribonucleoprotein A2 (HNRNPA2) (Shan et al. 2003), are involved in the trafficking of mRNAs within neurons. RBPs that bind in the CDS can also function in translation. This includes roles as both positive (Sanford et al. 2004; Peng et al. 2011) and negative (Darnell et al. 2011) regulators of translation, as well as in the regulation of alternative translation initiation site usage (Bonnal et al. 2005). As a final example, insulin like growth factor 2 mRNA binding protein 1 (IGF2BP1) has been found to stabilize some of its mRNA targets (Noubissi et al. 2006). Several of the RBPs alluded to in this paragraph or in the cited literature are indeed associated to motif sets that have positive normalized density (i.e. are enriched over expected) in intronless CDSs. However, because our method inherently comes with a certain amount of uncertainty with regards to the motif to RBP mapping, we prefer not to draw inferences with regards to the importance of any individual RBPs (see *Materials and Methods*).

2. Evidence that coding sequence evolution is constrained by the need to prevent inappropriate interactions with RBPs.

A novel result of this study is the finding that coding regions appear to be under selection to avoid certain RBP motifs. This is supported by evidence for selection against substitutions that would generate such a motif. We also predicted that when the presumed avoided motifs do occur, they would be evolving faster than random expectations, reflecting selection for degradation. We found no such evidence. This may suggest that the purifying selection to avoid the motifs is sufficiently efficient to prevent their fixation in locations where they might have a deleterious effect. Given the rarity of these motifs, however, it is also possible that we simply lack power to detect any increase in evolutionary rates.

This pattern of conserving certain regulatory sequences, yet selectively avoiding others, is likely not specific to RBP motifs but is rather a general feature of genome evolution. Indeed, there is evidence that the 3'UTRs of genes that are co-expressed with a microRNA are depleted in target sites to that microRNA, most likely to prevent inappropriate down-regulation (Bartel and Chen 2004; Farh et al. 2005; Stark et al. 2005; Chen and Rajewsky 2006), although see Iwama et al. (2007). Other examples of such avoidance selection include selection against spurious transcription factor binding sites in prokaryotes (Hahn et al. 2003) and in yeast (Babbitt 2010), as well as against mononucleotide runs within coding regions in various organisms, potentially to decrease the probability of transcriptional or translational error (Ackermann and Chao 2006; Gu et al. 2010; Itzkovitz et al. 2010). To our knowledge, the present work is the first large-scale study to consider selection to avoid RBP motifs.

Importantly, our results suggest that multiple coding between regulatory and protein structure information is not just about increased purifying selection at the locations where overlapping regulatory signals occur. It also places a more large-scale bias upon the sequence space available in coding region evolution. Not only are regions where necessary regulatory elements appear constrained not to lose them, all coding sequence is expected to be under some level of evolutionary constraint so as not to gain inappropriate signals. The latter constraint is likely to be weaker: given a functional motif, a large fraction of the possible mutations are expected to disrupt it, whereas a much more limited number of mutations would turn a non-motif into an (avoided) motif.

Finally, we would like to emphasize that our categorization of RBP motifs as preferred or avoided (or neither) is necessarily a gross simplification. Many relevant factors, which might help refine our crude approximations, have not been taken into account. For instance, we have not attempted to predict the mRNA secondary structure around motif hits. This could be relevant, as certain motifs may be preferred/avoided only when the site is accessible. Another important variable that is not considered is that of the context in which the motif hits appear. This includes both the sequence context – the other *k*-mers occurring in the vicinity – and the gene anatomic context, for instance, whether the site is located at an exon end or in the exon core. Some of the motif sets that currently appear to distribute and evolve according to chance expectations might turn out to show evidence of selection once such factors have been accounted for. However, analyses of this type have a great propensity to produce spurious patterns and so they should only be performed with explicit, well-motivated hypotheses in mind.

3. Future directions

Our results indicate that although the need to preserve RBP interactions has a detectable and significant impact on CDS evolution, the effect is slight (though, as touched upon in section 1 of the *Discussion*, the figures that we provide are likely underestimates). Studies on ESEs have reached similar conclusions (Parmley et al. 2006; Cáceres and Hurst 2013). However, these results appear, at first sight, to contradict a separate line of work where researchers have experimentally introduced large numbers of mutations into exons to determine the effect on splicing (Pagani et al. 2003; Pagani et al. 2005; Tournier et al. 2008; Gaildrat et al. 2012; Di Giacomo et al. 2013; Mueller et al. 2015; Julien et al. 2016; Soukarieh et al. 2016; Tajnik et al. 2016). Such studies have inferred an unexpectedly large proportion of exonic sites to be involved in splicing (over 90% according to the highest estimate (Julien et al. 2016)), suggesting that the need to maintain correct RNA processing could, on the contrary, be a major factor in CDS evolution. Are the results from these two independent fields of investigation comparable? Why do they appear to lead to such contrasting views on the prevalence and the evolutionary impact of exonic splice regulation (and of exonic RBP interactions more globally)? Finding answers to these questions will help us understand better the evolutionary dynamics of non-coding information within CDSs but might also shed light on other fundamental problems, such as estimating the extent to which variation in alternative splicing patterns is functional.

Materials and Methods

Caveats and methodological clarifications

The aim of the current work was to understand better how selection pressures related to RBP-binding have shaped human CDSs. It must be emphasized that our results are only indirectly relevant to the related problem of determining where on (pre-)mRNAs interactions with RBPs actually occur. Primary sequence is only one determinant of where an RBP binds, and can be more or less important depending on the protein (Li et al. 2014). For example, the binding preferences of many RBPs appear to be highly sensitive to local mRNA secondary structure (e.g. Wu et al. 2004; Aviv et al. 2006; Oberstrass et al. 2006; Skrisovska et al. 2007; Li et al. 2010; Masliah et al. 2013; Lambert et al. 2014). Because of this, our method, which consists solely in scanning the sequence for particular *k*-mers, cannot be used to determine individual binding sites with any accuracy. However, if the over-all density or rate of evolution of a set of motifs deviate from neutral expectations, this is likely an indication that selection has acted upon the motifs. It is precisely these kinds of patterns that we study and quantify in the current paper.

If one wishes to obtain a snapshot of the protein-RNA interactions occurring in a population of cells at a given time, approaches such as ours are inappropriate. One then typically turns to various genome-wide experimental methods based on the crosslinking and immunoprecipitation of protein-RNA complexes, followed by high-throughput sequencing of the RNA fragments (CLIP-seq) (Licatalosi et al. 2008; Xue et al. 2009; Hafner et al. 2010; Konig et al. 2011; Van Nostrand et al. 2016). Although caveats apply (Kishore et al. 2011; Sugimoto et al. 2012; Friedersdorf and Keene 2014; Lambert et al. 2014)), these methods are the state of the art for localizing RBP target sites on RNA.

However, data from CLIP-seq studies cannot easily be used to assess the long-term evolutionary impact of RBP-protein interactions, which is our goal in this paper. By its very nature, the method does not distinguish between spurious binding and evolutionarily relevant interactions − that a given interaction is observed, even if repeatably and significantly above background, does not always mean that it has fitness relevance or an impact on sequence evolution. In addition, CLIP-seq data does not allow one to precisely control for nucleotide composition biases, a crucial confound in any analysis of molecular evolution. Finally, producing estimates of global evolutionary impact is further rendered difficult by a high false negative rate (Darnell 2010; though see Van Nostrand et al. 2016).

Computational methods, such as the one used in this work, are therefore more appropriate for answering questions on sequence evolution. Several caveats must, nevertheless, be bourn in mind. Firstly, although the motifs used in this study were derived through experiments conducted on particular RBPs, there is nevertheless no direct link between motif and RBP during the sequence analysis. Similar motifs can be recognized by different RBPs (for instance, in our dataset, the motif *CCATACC* is associated to both poly(RC) binding protein 1 (PCBP1) and to heterogeneous nuclear ribonucleoprotein K (HNRNPK)). This means that when a set of motifs displays interesting distributional or evolutionary properties, there is no guarantee that this is necessarily due to interactions with the RBP to which we have associated that set of motifs, rather than to any other roles the motifs might have. We note that motif sets associated to RBPs that have been experimentally observed to preferentially bind within coding sequence are also at a greater density (raw and normalized) in coding regions than those predicted to bind elsewhere (section 3.1. in *Results*). This suggests that the motif to RBP mapping does indeed have global validity. However, it is still advisable to limit interpretation to over-all patterns (such as the relationship between enrichment and conservation measures) rather than to draw conclusions regarding particular RBPs.

In addition, the extent of sequence-specificity is expected to vary between RBPs (Li et al. 2014; Jankowsky and Harris 2015). Therefore, if a set of motifs associated to a particular RBP distributes in accordance with random expectations, this does not necessarily mean that interactions with this protein are unimportant for CDSs. It may simply indicate that *in vivo*, sequence is not a very important determinant of where this RBP binds. On a similar note, the quality of the motif sets is likely to vary depending on the protein and the method used to derive the motifs, with different techniques plagued by different biases (Marchese et al. 2016). This could also partially explain why certain sets of motifs show stronger deviations from neutrality than others.

General methods

The majority of the analysis was conducted using custom Python 3.4.2. and Perl v5.22.2 scripts (code available at www.github.com/rosinaSav/RBP_motifs). Unless otherwise noted, only standard libraries, NumPy 1.9.1. (van der Walt et al. 2011) and Biopython 1.64 (Cock et al. 2009) were used. R version 3.2.1. (R Core Team 2013) was used for plotting and for pre-made statistical tests. Bedtools 2.19.1 (Quinlan and Hall 2010) was used for operations on sequence coordinates. The analysis of human and macaque was based on assemblies GRCh38 and MMUL1, with the annotations corresponding to Ensembl release 78 for CDSs and Ensembl release 85 for non-coding regions (Cunningham et al. 2015). For the mouse and rat analysis, genome assemblies GRCm38 and Rnor_6.0 with the annotations from Ensembl release 80 were used. The genome sequences were obtained from the UCSC database (Karolchik et al. 2004). Gene annotations were downloaded as .gtf files from the Ensembl FTP site (Cunningham et al. (2015); ftp.ensembl.org/pub, last accessed 25 August 2015 for human (release 78) and mouse, 30 October 2015 for rat and 19 August 2016 for human (release 85) and macaque). Ensembl BioMart was used for retrieving the macaque CDS sequences (Kinsella et al. 2011; http://www.ensembl.org/biomart/martview; last accessed 21 February 2015). The pairwise alignments of human and macaque non-coding regions were retrieved from the Ensembl Compara database (Herrero et al. 2016) using a local installation of the Ensembl database and API (release 85).

The RBP motif sets

Consensus motifs for the various RBPs were retrieved from several sources, detailed below. Some sources store position weight matrices (PWMs) or position-specific scoring matrices (PSSMs), while others use consensus sequences. We converted the PWMs/PSSMs into consensus sequences by representing each site in the matrix as the IUPAC symbol corresponding to all those bases that presented a value greater than 0 (in the case of PWMs) or 0.25 (in the case of PSSMs) at that site.

RBPDB

The *all experiments* and *all proteins* CSV files were downloaded from rbpdb.ccbr.utoronto.ca/download.php (Cook et al. 2011; last accessed 11 November

2015). Those experiments that were not performed in *Homo sapiens*(*/Mus musculus*) or for *Homo sapiens*(*/Mus musculus*) RBPs, or that did not report a sequence motif were excluded. The consensus motifs from the remaining experiments were retained. In addition, PWMs were downloaded from the same website and converted into consensus sequences, as described above.

RBPmap

The RBPmap package was downloaded from rbpmap.technion.ac.il/download.html (Paz et al. 2014; last accessed 12 November 2015). PSSMs for human/mouse proteins were converted into consensus motifs. RBPmap does not distinguish between human and mouse and so the PSSMs retained for either analysis were identical, except that PSSMs originating from RNAcompete were ignored for mouse (this was to avoid including a large set of PSSMs determined originally for human in the mouse analysis).

SFmap

SFmap consensuses were obtained from sfmap.technion.ac.il/SF_list.html (Paz et al. 2010; last accessed 12 November 2015) and added to the list of motifs. SFmap does not distinguish between human and mouse and so all the motifs were included when analysing either species.

CISBP-RNA

The entire *Homo sapiens* dataset was retrieved from cisbp-rna.ccbr.utoronto.ca/bulk.php (Ray et al. 2013; last accessed 11 November 2015). The PSSMs labelled *direct* (signifying that the motifs were experimentally determined for that particular RBP rather than inferred from proteins with similar domains) were retained and converted into consensus sequences. Mouse consensuses were derived similarly, except that indirect PSSMs were also included.

Motifs from the different sources were then pooled. This resulted in 183 RBPs in human and 188 in mouse, each associated to a set of *k*-mers. *N* (fully ambiguous) bases at the very beginning or at the very end of motifs were removed. The motifs were then filtered to only leave those with length between 5 and 12 bases (included). Motifs that contained parentheses (signifying variable motif length) were removed.

After this filtering step, 133 RBPs remained in human and 163 in mouse. However, because the source databases differed in naming conventions, some of the RBP identifiers that had been retained referred to the same protein. For human, the remaining RBP identifiers were therefore fed to Ensembl BioMart and converted to Ensembl gene identifiers. This step was undertaken to verify whether or not the identifiers were recognized as valid HGNC symbols. Those that were not were manually converted into HGNC symbols using the GeneCards database (www.genecards.org (Safran et al. 2010); last accessed 12 November 2015). For mouse, the protein identifiers were input into the Mouse Genome Informatics (MGI (Bult et al. 2016); last accessed 19 October 2016) web site as a batch query. The output was used to update all identifiers to the *current symbol* recognized for the protein. Hnrnpcl1, which was not recognized at all by MGI, was discarded. This step resulted in several synonymous identifiers being collapsed, leaving us with a total of 117 RBPs for human and 81 for mouse. Three of the human RBPs were removed from the dataset: microRNA 1236 (*MIR1236;* because it is a microRNA gene rather than an RBP), poly(A) binding protein, cytoplasmic 4 (PABPC4; the consensus was *AAAAAAA,* making normalization for dinucleotide composition impossible) and peptidylprolyl isomerase E (cyclophilin E) (PPIE; the consensus was *WWWWWW,* making it once again impossible to generate simulants). Pabpc4 was removed from the mouse set for similar reasons. The final number of RBPs retained was therefore 114 for human and 80 for mouse.

In human, two consensus sequences were added manually: the consensus *UUWGDUU* was added to ELAV like RNA binding protein 1 (ELAVL1), while the consensus *RWUUYAUUUWR* was added to ELAV like neuron-specific RNA binding protein 2 (ELAVL2). This is because in these cases, the retained motifs included both consensus sequences lifted directly from a database, as well as consensuses that we had derived from a PWM/PSSM. Both, however, were based on the same original publication. The new motifs were added to summarize these existing consensuses in a broader consensus that would combine the information from both sources.

For all RBPs, the remaining consensuses were then expanded into all the non-ambiguous motifs that would match the consensus. Identical motifs were collapsed. This resulted in the final motif sets (Additional File 1).

The random motifs (used only to generate the distribution indicated by a black line in Figure 1A)

The sequence of the human genome (GRCh38) was obtained from the UCSC Genome Browser website (Karolchik et al. 2004). Only reference chromosomes were considered: unplaced, unlocalized and alternative sequences were excluded. The counts of each of the 4 DNA bases were summed across all the chromosomes and divided by the total number of canonical (*A*, *T*, *C* or *G*) bases.

In each of the RBP motif sets, the motifs were then replaced by random motifs of the same length. To generate a motif of length *k*, *k* canonical bases were randomly picked (using *numpy.random.choice()*), with the probability of each base being chosen corresponding to its mononucleotide frequency in the human genome, as determined above.

The sequence sets

Full CDSs

The sets of intron-containing and intronless human CDS sequences were the same as those used in Savisaar and Hurst (2016). The methods used for generating these sequence sets were detailed in the cited publication and will only briefly be summarized here. All intronless/intron-containing ORFs from *GRCh38* were downloaded from the Ensembl database (release 78). For intronless genes, only ORFs from genes that exclusively produced intronless transcripts (according to the transcript annotations available in the Ensembl database) were kept. The ORFs were then checked for reading frame integrity and completeness. If several transcripts corresponded to one gene, the one with the longest ORF was kept. The remaining transcripts were then aligned to macaque orthologs. Only those that had an ortholog to which they aligned with a d_S below 0.2 and a d_N/d_S below 0.5 were kept. This filtering step was necessary to minimize the proportion of pseudo-genes in the set. Finally, the sequences were BLASTed all against all and clustered into paralogous families based on the results. Mouse full CDSs were obtained similarly (using *GRCm38*, Ensembl release 80), except that the d_S threshold was set to 0.3 during the filtering.

Exon flanks and cores

To generate the sets of exon flanks/cores, we recovered all of the internal fully coding exons in our set of human intron-containing genes that were at least 211 base pairs long (based on Ensembl release 78 annotations; only one randomly picked gene was considered per paralogous family). The exons were trimmed so as to both start and end with full codons. (The length threshold was set to 211 because three 69 base pair long non-overlapping regions were to be extracted from each exon (3 * 69 = 207) and at least 4 base pairs had to be left over in case any nucleotides were lost because of trimming.) Three sequence regions were then extracted: the first 69 base pairs at the 5' end of the exon, the final 69 base pairs at the 3' end and 69 base pairs from the very centre. If the number of codons separating the 5' region from the 3' region was even, meaning that it was not possible to define the exact mid-point (when 69 is subtracted from an even number, the result is odd), the core was defined so as to be separated from the 5' flank by n codons and from the 3' flank by $n - 1$ codons.

Non-coding sequences

A set of human CDSs was retrieved and filtered for ORF integrity and conservation level as per the procedure used above, except that Ensembl release 85 annotations were used. The sequences were clustered into putative paralogous families as described above. The chromosomal coordinates of the introns, 5'UTRs and 3'UTRs associated to the transcripts in the set were retrieved based on Ensembl gene annotations (release 85, one transcript was randomly picked from each paralogous family). In addition, 100 base pairs were extracted from immediately upstream and immediately downstream of each exon (only the intronic flank was used for terminal exons). The full introns set was filtered further, firstly by randomly picking only one intron from each transcript (to limit the size of the dataset for computational reasons), and secondly by excluding all introns that overlapped with any exons, as defined by Ensembl annotations.

The coordinates were then used to retrieve the LASTZ_NET human-macaque pairwise alignment from a local installation of the Ensembl Compara database (release 85), using the Ensembl API. Only alignments that corresponded to a single genome alignment block and that contained no *N* bases in either the human or the macaque sequence were retained.

Motif density and ND

To calculate the density of the full set of motifs, we counted the number of bases that overlapped with any of the motifs in the set in each CDS and divided this count by the length of the CDS. We used the full CDS, that is to say, all of the coding sequence between the start and the stop codon in the relevant transcript variant. We did not take into consideration the positions of exon-exon junctions. Bases encompassed by more than one motif were only counted once (i.e. overlapping motifs were collapsed). We calculated an ND value separately for each gene (see main text for the calculation of ND and below for the generation of simulant motifs) and used the median density and the median ND as our statistics (averaged across paralogous families). Because less data was available, a different approach was used when calculating the densities of individual motif sets (the motifs associated to a particular RBP). Namely, rather than producing a density estimate per gene, we summed the number of overlapping bases across all the sequences and divided that by the summed length of the sequences. This produced a single point estimate for density and for ND for each set of motifs. Counts and lengths were averaged across paralogous families before the division step.

1000 (for the full set density analysis in human CDSs) or 100 (for the full set density analyses in human non-coding sequences and for the mouse analysis) simulant versions of the RBP motifs set were generated in order to calculate the enrichment *p*-value and ND. The motifs were divided into dinucleotides in the two possible phases. To generate each of the motifs in the 1000/100 simulant sets, the necessary number of dinucleotides were sampled randomly with replacement from the pool of dinucleotides. If the motif length was odd, an additional base was sampled from the mononucleotide composition of the motif set. This resulted in 1000/100 sets of simulant motifs, with the motif number, motif lengths and the dinucleotide composition matched to the true set of motifs (the match being approximate in the case of dinucleotide composition). The resulting simulants were screened, such that no simulated motifs were allowed that also appeared in the set of real motifs. In addition, no simulants could contain a mononucleotide run that was longer than the longest run of that base in the real motif set. Finally, all the motifs within a particular simulant set had to be unique. In the analysis of motif enrichment independent of stop codon content, simulants were additionally constrained to be devoid of the substrings *TAA*, *TGA* and *TAG* (see Supplementary Text 3 in Additional File 5). Simulants were generated similarly for the individual motif sets (1000 simulant sets were always used).

Hits were then predicted in the sequences to each of the simulated motif sets, generating an empirical distribution of simulated density values. From this distribution, ND and *p* were derived as described in the main text (see above for differences between the processing of the full motifs set and the individual sets). The normalization step is even more important when considering individual sets of motifs (motifs grouped based on the putative cognate RBP), as in addition to controlling for nucleotide composition biases, this step largely eliminates the confounding factor of the sets varying in the number and length of the motifs. For instance, the smallest sets only consist of a single motif whereas the largest in human − composed of *k*-mers putatively recognized by the RBP transformer-2 protein homolog beta (TRA2B) − has 218 motifs.

After calculating the density of the individual motif sets, we noticed that some were very rare, leading to concerns over whether there was sufficient information to reliably estimate ND and other parameters in those cases. In human, we decided to only include those motif sets in the subsequent analysis that filled one of two criteria: either the hits to the real motifs totalled at least 100 bp in the intron-containing CDSs, as well as in each of four other sequence sets (intronless CDSs, exon 5' flanks, exon cores and exon 3' flanks) or the hits to at least half of the simulant sets did. The reasoning behind this rule was that if the real motifs were rare, whereas the simulants were not, or the other way around, then this was potentially a biologically meaningful pattern, whereas if both were rare, then one simply had a lack of information. The mouse filtering was similar, except that only the density in intron-containing CDSs was considered. Only one RBP was filtered out in this process (Raver1).

Rate of evolution at synonymous sites

d_S estimates were calculated identically to Savisaar and Hurst (2016), which details the methods used. Only a brief summary will therefore be provided here. Sequence regions overlapping with RBP motifs were extracted and aligned to homologous regions in macaque (*Macaca mulatta*). The rate of evolution at synonymous sites was calculated using the Goldman and Yang (Goldman and Yang 1994) method, as implemented in the *codeml* programme that is part of the Phylogenetic Analysis by Maximum Likelihood (PAML) (Yang 2007) suite. This procedure was then repeated for each of 1000 simulant sets, enabling us to calculate a normalized d_S estimate and an enrichment *p*-value. One randomly picked gene was considered from each paralogous family.

Rate of evolution of non-coding sequence

To calculate rates of evolution for non-coding sequences, the *baseml* programme from the PAML suite was used (*model* = 1). The statistic used, termed here d_{NC}, corresponds to the tree length reported by the programme.

Conservation at four-fold degenerate sites overlapping different dinucleotides

The four-fold degenerate sites in intron-containing sequences were divided into two groups: those that overlapped an RBP motif hit and those that did not. Within each class, we then further sub-divided the sites based on the overlapping dinucleotide. Each site was counted twice, once as belonging to the dinucleotide in which it was the second base and once as belonging to the dinucleotide in which it was the first base. For each dinucleotide class within either site type (motif or non-motif), we determined the fraction of sites where the orthologous position in macaque did not exhibit the same base as in human. In order to obtain an over-all estimate of the difference in evolutionary rate between motif and non-motif, we averaged the rates calculated for different dinucleotides but weighted them by the frequency of each dinucleotide within the subset of sites overlapping with RBP motifs, thereby controlling for any differences in dinucleotide composition between motif and non-motif regions. A random member was included fro each paralogous family.

Human-macaque comparison at four-fold degenerate sites that are a single base substitution away from a putatively avoided motif in human

We determined all four-fold degenerate sites in our set of full human intron-containing CDSs (one randomly picked gene from each paralogous family) such that a single base substitution at that site could generate a putatively avoided motif (a motif with enrichment *p*-value above 0.9 in full intron-containing CDSs). We then scored each site based on the identity of the orthologous macaque base. The following scores were possible: 0 (the base present in macaque is either identical to that present in human or is a base other than the one(s) that would give rise to an avoided motif in human), 0.25 (the base present in macaque would give rise to an avoided motif in human. Of the 3 possible base substitutions in human, all three would generate a putatively avoided motif), 0.5 (the base present in macaque would give rise to an avoided motif in human. Of the 3 possible base substitutions in human, two would generate a putatively avoided motif) and 0.75 (the base present in macaque would give rise to an avoided motif in human. Of the 3 possible base substitutions in human, only the one used in macaque would generate a putatively avoided motif). The scores were summed across all sites and the sum divided by the number of sites considered. The analysis was then repeated on 1000 sets of simulated motifs that broadly matched the dinucleotide composition of the putatively avoided motifs, allowing us to calculate a *p*-value for the statistic obtained.

The reasoning behind the scoring system is that macaque presenting the base that would generate the avoided motif in human constitutes stronger evidence against avoidance against that motif if other substitutions were possible that would not have generated the motif than when any substitution would have led to a putatively avoided motif. Note that there are several caveats to this analysis. Firstly, because we did not use an outgroup, we do not know on which the branch the substitution occurred in cases where the human and the macaque sequence differ. However, more frequently than expected by chance, macaque also does not have the base that would give rise to the putatively avoided motif in human, suggesting that this was also the case in the most recent common ancestor. Secondly, it is possible that a substitution that would generate a particular putatively avoided motif would simultaneously disrupt another such motif that overlaps with the first, meaning that the substitution would not necessarily lead to an increase in avoided motif density. Our analysis did not consider this issue. Thirdly, in macaque, we only analysed the base present at the particular site considered. We therefore did not account for any other potential differences between human and macaque at sites nearby, which could mean that even though a particular substitution would lead to a putatively avoided motif in human, it might not do so in macaque.

Annotating the motif sets based on the properties of the associated RBP

To annotate the motif sets based on the binding profile of the associated RBP, we searched the literature for high-throughput crosslinking and immunoprecipitation (CLIP-seq) studies conducted on that RBP. Only one study was considered per RBP. Each RBP was annotated as either *CDS* or *other* based on whether or not the study reported an enrichment of binding clusters in the CDS (if no CLIP-seq studies could be found, the RBP was annotated as *NA*). The interpretation of the authors was followed when deciding how to report the results of a particular study. For instance, if the authors reported CDS clusters to be rare but did not control for the fact that the combined length of coding regions is much shorter than that of introns, we still annotated the RBP as *other*. The annotations, as well as the sources used, are listed in Supplementary Spread Sheet 5 in Additional File 4.

Expression analysis

The phase 1 and 2 combined normalized .osc file was retrieved from the FANTOM5 website (http://fantom.gsc.riken.jp/5/datafiles (Fantom Consortium et al. 2014); last accessed 11 February 2016). The data was filtered to only leave samples where the sample name contained the substring *adult, pool1*. All brain tissues except for the full brain sample and the retinal sample were removed. Peak coordinates were converted to *hg38* coordinates using CrossMap 0.2.2. (Zhao et al. 2014). For each transcript in our set of intron-containing protein-coding genes (based on Ensembl release 78), we defined a region of 1001 base pairs centered on the start coordinate of the Ensembl transcript annotation as the promoter and associated all peaks that overlapped that promoter to that peak. If several peaks were associated to a single transcript, we summed the tags per million (TPM) within each sample across the peaks. The TPM were then averaged across paralogous families.

Acknowledgements

The authors would like to thank Nizar Batada for helpful discussions. This work was supported by the Boehringer Ingelheim Fonds (PhD fellowship to R.S.), the European Research Council (Advanced grant ERC-2014-ADG 669207 to L.D.H.), and the Medical Research Council (MR/L007215/1 to L.D.H.).

Tables

Table 1: Spearman correlation between normalized density (ND) and various expression parameters, determined based on FANTOM5 data.

	expression breadth (fraction of tissues where gene is expressed[a])	maximum expression	median expression	median expression in tissues where the gene is expressed
ρ	≈–0.151	≈–0.035	≈–0.157	≈–0.016
p[b]	≈$9.576*10^{-30}$ (≈$3.830*10^{-29}$)	≈0.010 (≈0.038)	≈$3.071*10^{-32}$ (≈$1.228*10^{-31}$)	≈0.280 (1.000)

[a] A gene is considered to be expressed in a given tissue if more than 5 tags per million map to the promoter region (see *Materials and Methods* for further details).

[b] The parentheses contain the Bonferroni-corrected p-value.

Table 2: Motif density and conservation parameters for various genic regions.

	CDSs	5'UTRs	3'UTRs	introns	upstream intronic[a]	downstream intronic[a]
median motif density	≈0.573	≈0.537	≈0.573	≈0.578	≈0.580	≈0.560
median ND[b]	≈0.115	≈0.145	≈0.103	≈0.129	≈0.167	≈0.130
enrichment p[c]	≈0.001	≈0.010	≈0.010	≈0.010	≈0.010	≈0.010
d_{NC}[d] or d_S[e]	≈0.064	≈0.052	≈0.043	≈0.055	≈0.051	≈0.054
normalized d_{NC}[d] or normalized d_S[e]	≈–0.041	≈–0.019	≈–0.026	≈–0.034	≈–0.035	≈–0.017
conservation p[c]	≈0.003	≈0.030	≈0.040	≈0.010	≈0.030	≈0.149
global reduction[f]	≈–2.4%	≈–1.0%	≈–1.5%	≈–2.0%	≈–2.0%	≈–0.9%

[a] Upstream/downstream intronic regions correspond to 100 bp slices immediately upstream/downstream from an exon.

[b] normalized density

[c] One-tailed p derived from an empirical distribution of simulant statistics. 1000 simulants were used for CDSs and 100 in the other cases.

[d] rate of evolution at non-coding sites. Used for all sequence regions except for CDSs.

[e] rate of evolution at synonymous sites. Used for CDSs.

[f] The global reduction is the product of the motif density and the conservation statistic (multiplied by 100). It is an estimate for the extent to which the (synonymous) substitution rate is decreased in the relevant region because of selection to preserve RBP motifs. Note that in the table, the density and the normalized conservation estimates have been rounded to the third decimal, whereas exact values were used when calculating the global reduction.

Figures

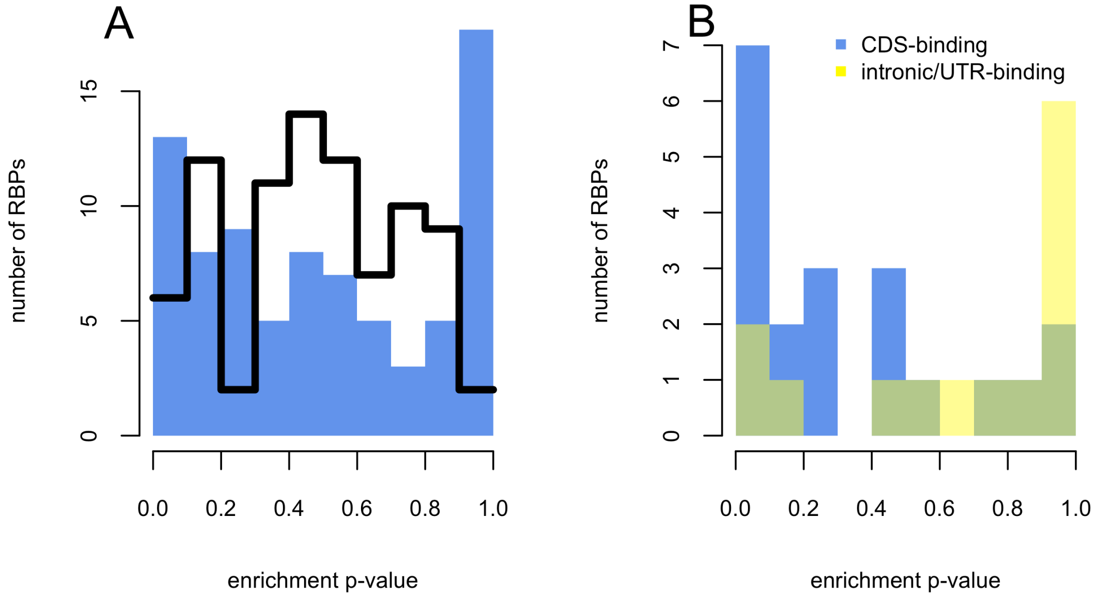

Figure 1: A. Each data point corresponds to the probability that a given motif set (recognized by a particular RBP) would be found at its current density (or higher) by chance given the underlying dinucleotide composition. The black line traces the distribution of enrichment *p*-values obtained in the same sequences for size-matched sets of random *k*-mers. Note that RBP motifs display a peak at either extreme of the distribution whereas the random motifs do not. In other words, RBP motifs show a disproportionate tendency to occur at a density that deviates from neutral expectations. Importantly, this can mean both enrichment (*p*-value approaching 0) and depletion (*p*-value approaching 1). B. As A, except that only RBPs for which we found crosslinking and immunoprecipitation studies on binding preferences are shown. Motif sets associated to CDS-binding RBPs (blue) have a peak near 0 (enrichment), whereas the other sets (yellow) have a peak near 1 (depletion).

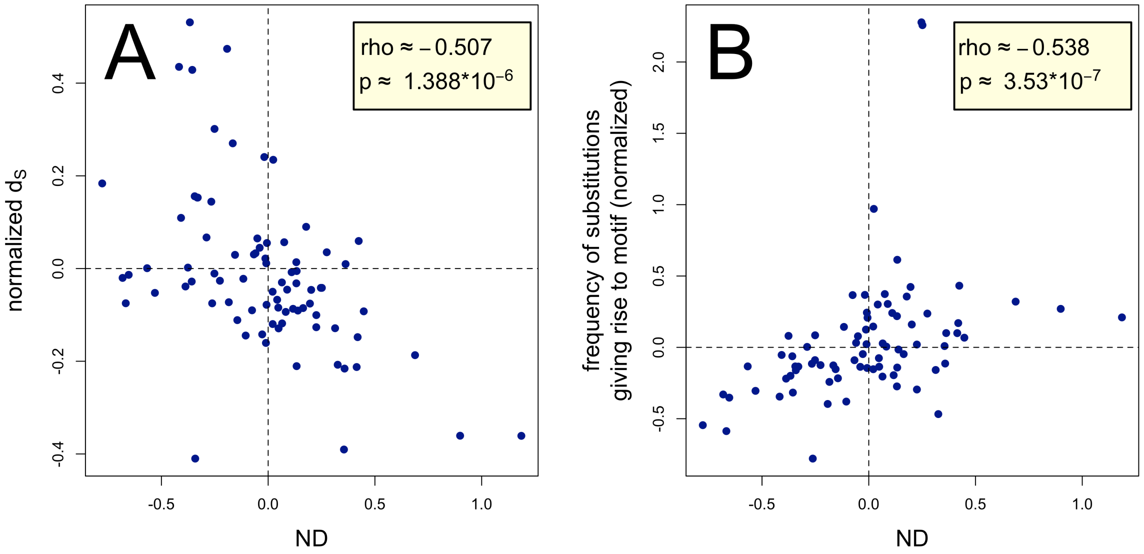

Figure 2: A. Correlation between a motif set's normalized density (ND) and its nucleotide-normalized d_S from alignment to macaque. Motif sets that are more strongly enriched are also more conserved, controlling for nucleotide composition. The dashed lines intersect the plot at the points where expected and observed frequencies would be equal. B. Correlation between ND and the nucleotide-normalized propensity to gain the motifs over evolution (measured by determining how frequently macaque sites that are orthologous to human fourfold degenerate sites that are a single base substitution away from the motif in human contain the base that would give rise to the motif in human). Note that because our analysis did not make use of an outgroup, we cannot know on which branch the substitution occurred in cases where the human and macaque sequence differ. See caption to subplot A for interpretation of the dashed lines.

References

UCSC Genome Browser Home [Internet]. [accessed 2015 13 Nov]. Available from: http://genome.ucsc.edu/

Ackermann M, Chao L. 2006. DNA sequences shaped by selection for stability. PLoS Genet 2:e22.

Agoglia RM, Fraser HB. 2016. Disentangling Sources of Selection on Exonic Transcriptional Enhancers. Mol Biol Evol 33:585-590.

Änkö M-L, Müller-McNicoll M, Brandl H, Curk T, Gorup C, Henry I, Ule J, Neugebauer KM. 2012. The RNA-binding landscapes of two SR proteins reveal unique functions and binding to diverse RNA classes. Genome biology 13:R17.

Ascano M, Jr., Mukherjee N, Bandaru P, Miller JB, Nusbaum JD, Corcoran DL, Langlois C, Munschauer M, Dewell S, Hafner M et al. 2012. FMRP targets distinct mRNA sequence elements to regulate protein expression. Nature 492:382-386.

Aviv T, Lin Z, Ben-Ari G, Smibert CA, Sicheri F. 2006. Sequence-specific recognition of RNA hairpins by the SAM domain of Vts1p. Nat Struct Mol Biol 13:168-176.

Babbitt GA. 2010. Relaxed selection against accidental binding of transcription factors with conserved chromatin contexts. Gene 466:43-48.

Barrell BG, Air GM, Hutchison III CA. 1976. Overlapping genes in bacteriophage φX174. Nature 264:34-41.

Bartel DP, Chen C-Z. 2004. Micromanagers of gene expression: the potentially widespread influence of metazoan microRNAs. Nature Reviews Genetics 5:396–400.

Belshaw R, Pybus OG, Rambaut A. 2007. The evolution of genome compression and genomic novelty in RNA viruses. Genome Res 17:1496-1504.

Birnbaum RY, Patwardhan RP, Kim MJ, Findlay GM, Martin B, Zhao J, Bell RJ, Smith RP, Ku AA, Shendure J et al. 2014. Systematic dissection of coding exons at single nucleotide resolution supports an additional role in cell-specific transcriptional regulation. PLoS Genet 10:e1004592.

Blencowe BJ. 2000. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. Trends in Biochemical Sciences 25:106–110.

Bonnal S, Pileur F, Orsini C, Parker F, Pujol F, Prats AC, Vagner S. 2005. Heterogeneous nuclear ribonucleoprotein A1 is a novel internal ribosome entry site trans-acting factor that modulates alternative initiation of translation of the fibroblast growth factor 2 mRNA. J Biol Chem 280:4144-4153.

Bult CJ, Eppig JT, Blake JA, Kadin JA, Richardson JE, Mouse Genome Database G. 2016. Mouse genome database 2016. Nucleic Acids Res 44:D840-847.

Cáceres EF, Hurst LD. 2013. The evolution, impact and properties of exonic splice enhancers. Genome biology 14:1–18.

Carlini DB, Genut JE. 2006. Synonymous SNPs provide evidence for selective constraint on human exonic splicing enhancers. J Mol Evol 62:89-98.

Chamary JV, Hurst LD. 2005. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. Genome biology 6:R75.

Chen K, Rajewsky N. 2006. Natural selection on human microRNA binding sites inferred from SNP data. Nat Genet 38:1452-1456.

Chirico N, Vianelli A, Belshaw R. 2010. Why genes overlap in viruses. Proc Biol Sci 277:3809-3817.

Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25:1422-1423.

Cook KB, Kazan H, Zuberi K, Morris Q, Hughes TR. 2011. RBPDB: a database of RNA-binding specificities. Nucleic Acids Res 39:D301-308.

Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S et al. 2015. Ensembl 2015. Nucleic Acids Res 43:D662-669.

Darnell JC, Van Driesche SJ, Zhang C, Hung KYS, Mele A, Fraser CE, Stone EF, Chen C, Fak JJ, Chi SW. 2011. FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. Cell 146:247-261.

Darnell RB. 2010. HITS-CLIP: panoramic views of protein-RNA regulation in living cells. Wiley Interdisciplinary Reviews RNA 1:266-286.

Di Giacomo D, Gaildrat P, Abuli A, Abdat J, Frebourg T, Tosi M, Martins A. 2013. Functional analysis of a large set of BRCA2 exon 7 variants highlights the predictive value of hexamer scores in detecting alterations of exonic splicing regulatory elements. Hum Mutat 34:1547-1557.

Fairbrother WG, Holste D, Burge CB, Sharp PA. 2004. Single nucleotide polymorphism-based validation of exonic splicing enhancers. PLoS Biol 2:E268.

Falanga A, Stojanovic O, Kiffer-Moreira T, Pinto S, Millan JL, Vlahovicek K, Baralle M. 2014. Exonic splicing signals impose constraints upon the evolution of enzymatic activity. Nucleic Acids Res 42:5790-5798.

Fang Z, Rajewski N. 2011. The Impact of miRNA Target Sites in Coding Sequences and in 3'UTRs. PloS one 6.

Fantom Consortium, the RP, Clst, Forrest AR, Kawaji H, Rehli M, Baillie JK, de Hoon MJ, Haberle V, Lassmann T et al. 2014. A promoter-level mammalian expression atlas. Nature 507:462-470.

Farh KKH, Grimson A, Jan C, Lewis BP, Johnston WK, Lim LP, Burge CB, Bartel DP. 2005. The Widespread Impact of Mammalian MicroRNAs on mRNA Repression and Evolution. Science 310:1817-1821.

Forman JJ, Legesse-Miller A, Coller HA. 2008. A search for conserved sequences in coding regions reveals that the let-7 microRNA targets Dicer within its coding sequence. Proc Natl Acad Sci U S A 105:14879-14884.

Friedersdorf MB, Keene JD. 2014. Advancing the functional utility of PAR-CLIP by quantifying background binding to mRNAs and lncRNAs. Genome Biol 15.

Gaildrat P, Krieger S, Di Giacomo D, Abdat J, Revillion F, Caputo S, Vaur D, Jamard E, Bohers E, Ledemeney D et al. 2012. Multiple sequence variants of BRCA2 exon 7 alter splicing regulation. J Med Genet 49:609-617.

Garcia-Mayoral MF, Diaz-Moreno I, Hollingworth D, Ramos A. 2008. The sequence selectivity of KSRP explains its flexibility in the recognition of the RNA targets. Nucleic Acids Res 36:5290-5296.

Glisovic T, Bachorik JL, Yong J, Dreyfuss G. 2008. RNA-binding proteins and post-transcriptional gene regulation. FEBS Lett 582:1977-1986.

Goldman N, Yang Z. 1994. A Codon-based Model of Nucleotide Substitution for Protein-coding DNA Sequences. Mol Biol Evol 11:725-736.

Grellscheid S, Dalgliesh C, Storbeck M, Best A, Liu Y, Jakubik M, Mende Y, Ehrmann I, Curk T, Rossbach K et al. 2011. Identification of evolutionarily conserved exons as regulated targets for the splicing activator tra2beta in development. PLoS Genet 7:e1002390.

Gu T, Tan S, Gou X, Araki H, Tian D. 2010. Avoidance of long mononucleotide repeats in codon pair usage. Genetics 186:1077-1084.

Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, Ascano M, Jungkamp AC, Munschauer M et al. 2010. PAR-CliP--a method to identify transcriptome-wide the binding sites of RNA binding proteins. J Vis Exp 41.

Hahn MW, Stajich JE, Wray GA. 2003. The effects of selection against spurious transcription factor binding sites. Mol Biol Evol 20:901-906.

Hausser J, Syed AP, Bilen B, Zavolan M. 2013. Analysis of CDS-located miRNA target sites suggests that they can effectively inhibit translation. Genome Res 23:604-615.

Herrero J, Muffato M, Beal K, Fitzgerald S, Gordon L, Pignatelli M, Vilella AJ, Searle SM, Amode R, Brent S et al. 2016. Ensembl comparative genomics resources. Database (Oxford) 2016.

Huang Y, Gattoni R, Stévenin J, Steitz JA. 2003. SR Splicing Factors Serve as Adapter Proteins for TAP-Dependent mRNA Export. Mol Cell 11:837-843.

Huang Y, Steitz JA. 2001. Splicing Factors SRp20 and 9G8 Promote the Nucleocytoplasmic Export of mRNA. Mol Cell 7:899-905.

Hurst LD. 2006. Preliminary Assessment of the Impact of MicroRNA-Mediated Regulation on Coding Sequence Evolution in Mammals. Journal of Molecular Evolution 63:174-182 %U http://link.springer.com/110.1007/s00239-00005-00273-00232.

Ince-Dunn G, Okano HJ, Jensen KB, Park WY, Zhong R, Ule J, Mele A, Fak JJ, Yang C, Zhang C et al. 2012. Neuronal Elav-like (Hu) proteins regulate RNA splicing and abundance to control glutamate levels and neuronal excitability. Neuron 75:1067-1080.

Itzkovitz S, Hodis E, Segal E. 2010. Overlapping codes within protein-coding sequences. Genome Res 20:1582-1589.

Iwama H, Masaki T, Kuriyama S. 2007. Abundance of microRNA target motifs in the 3'-UTRs of 20527 human genes. FEBS Lett 581:1805-1810.

Jankowsky E, Harris ME. 2015. Specificity and nonspecificity in RNA-protein interactions. Nat Rev Mol Cell Biol 16:533-544.

Julien P, Minana B, Baeza-Centurion P, Valcarcel J, Lehner B. 2016. The complete local genotype-phenotype landscape for the alternative splicing of a human exon. Nat Commun 7:11558.

Kao D-I, Aldridge GM, Weiler IJ, Greenough WT. 2010. Altered mRNA transport, docking, and protein translation in neurons lacking fragile X mental retardation protein. Proceedings of the National Academy of Sciences 107:15601-15606.

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. 2004. The UCSC Table Browser data retrieval tool. Nucleic Acids Res 32:D493-496.

Ke S, Zhang XH, Chasin LA. 2008. Positive selection acting on splicing motifs reflects compensatory evolution. Genome Res 18:533-543.

Kinsella RJ, Kahari A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D, Derwent P, Kerhornou A et al. 2011. Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database (Oxford) 2011:bar030.

Kishore S, Jaskiewicz L, Burger L, Hausser J, Khorshid M, Zavolan M. 2011. A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. Nat Methods 8:559-564.

Konig J, Zarnack K, Rot G, Curk T, Kayikci M, Zupan B, Turner DJ, Luscombe NM, Ule J. 2011. iCLIP--transcriptome-wide mapping of protein-RNA interactions with individual nucleotide resolution. J Vis Exp 50.

Lambert N, Robertson A, Jangi M, McGeary S, Sharp PA, Burge CB. 2014. RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. Mol Cell 54:887-900.

Lewis BP, Burge CB, Bartel DP. 2005. Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets. Cell 120:15-20.

Li X, Kazan H, Lipshitz HD, Morris QD. 2014. Finding the target sites of RNA-binding proteins. Wiley Interdiscip Rev RNA 5:111-130.

Li X, Manley JL. 2005. Inactivation of the SR protein splicing factor ASF/SF2 results in genomic instability. Cell 122:365-378.

Li X, Quon G, Lipshitz HD, Morris Q. 2010. Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure. RNA 16:1096-1107.

Licatalosi DD, Mele A, Fak JJ, Ule J, Kayikci M, Chi SW, Clark TA, Schweitzer AC, Blume JE, Wang X et al. 2008. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature 456:464-469.

Lin MF, Kheradpour P, Washietl S, Parker BJ, Pedersen JS, Kellis M. 2011. Locating protein-coding sequences under selection for additional, overlapping functions in 29 mammalian genomes. Genome Research 21:1916-1928.

Liu G, Zhang R, Xu J, Wu CI, Lu X. 2015. Functional conservation of both CDS- and 3'-UTR-located microRNA binding sites between species. Mol Biol Evol 32:623-628.

Macossay-Castillo M, Kosol S, Tompa P, Pancsa R. 2014. Synonymous constraint elements show a tendency to encode intrinsically disordered protein segments. PLoS Comput Biol 10:e1003607.

Marchese D, de Groot NS, Lorenzo Gotor N, Livi CM, Tartaglia GG. 2016. Advances in the characterization of RNA-binding proteins. Wiley Interdiscip Rev RNA 7:793-810.

Masliah G, Barraud P, Allain FH. 2013. RNA recognition by double-stranded RNA binding domains: a matter of shape and sequence. Cell Mol Life Sci 70:1875-1895.

Meyer IM, Miklos I. 2005. Statistical evidence for conserved, local secondary structure in the coding regions of eukaryotic mRNAs and pre-mRNAs. Nucleic Acids Res 33:6338-6348.

Michel AM, Choudhury KR, Firth AE, Ingolia NT, Atkins JF, Baranov PV. 2012. Observation of dually decoded regions of the human genome using ribosome profiling data. Genome Res 22:2219-2229.

Mueller WF, Larsen LS, Garibaldi A, Hatfield GW, Hertel KJ. 2015. The Silent Sway of Splicing by Synonymous Substitutions. J Biol Chem 290:27700-27711.

Muller-McNicoll M, Neugebauer KM. 2013. How cells get the message: dynamic assembly and function of mRNA-protein complexes. Nat Rev Genet 14:275-287.

Normark S, Bergström S, Edlund T, Grundström T, Jaurin B, Lindberg FP, Olsson O. 1983. Overlapping genes. Annu Rev Genet 17:499-525.

Noubissi FK, Elcheva I, Bhatia N, Shakoori A, Ougolkov A, Liu J, Minamoto T, Ross J, Fuchs SY, Spiegelman VS. 2006. CRD-BP mediates stabilization of betaTrCP1 and c-myc mRNA in response to beta-catenin signalling. Nature 441:898-901.

Oberstrass FC, Lee A, Stefl R, Janis M, Chanfreau G, Allain FH. 2006. Shape-specific recognition in the structure of the Vts1p SAM domain with RNA. Nat Struct Mol Biol 13:160-167.

Pagani F, Buratti E, Stuani C, Baralle FE. 2003. Missense, nonsense, and neutral mutations define juxtaposed regulatory elements of splicing in cystic fibrosis transmembrane regulator exon 9. J Biol Chem 278:26580-26588.

Pagani F, Raponi M, Baralle FE. 2005. Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. Proc Natl Acad Sci U S A 102:6368-6372.

Pancsa R, Tompa P. 2016. Coding Regions of Intrinsic Disorder Accommodate Parallel Functions. Trends Biochem Sci.

Parmley JL, Chamary JV, Hurst LD. 2006. Evidence for Purifying Selection Against Synonymous Mutations in Mammalian Exonic Splicing Enhancers. Molecular biology and evolution 23:301-309.

Parmley JL, Urrutia AO, Potrzebowski L, Kaessmann H, Hurst LD. 2007. Splicing and the evolution of proteins in mammals. PLoS biology 5:e14.

Paz I, Akerman M, Dror I, Kosti I, Mandel-Gutfreund Y. 2010. SFmap: a web server for motif analysis and prediction of splicing factor binding sites. Nucleic Acids Res 38:W281-285.

Paz I, Kosti I, Ares M, Jr., Cline M, Mandel-Gutfreund Y. 2014. RBPmap: a web server for mapping binding sites of RNA-binding proteins. Nucleic Acids Res 42:W361-367.

Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, Lander ES, Kent J, Miller W, Haussler D. 2006. Identification and classification of conserved RNA secondary structures in the human genome. PLoS Comput Biol 2:e33.

Peng S, Chen LL, Lei XX, Yang L, Lin H, Carmichael GG, Huang Y. 2011. Genome-wide studies reveal that Lin28 enhances the translation of genes important for growth and survival of human embryonic stem cells. Stem Cells 29:496-504.

Pozzoli U, Riva L, Menozzi G, Cagliani R, Comi GP, Bresolin N, Giorda R, Sironi M. 2004. Over-representation of exonic splicing enhancers in human intronless genes suggests multiple functions in mRNA processing. Biochem Biophys Res Commun 322:470-476.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841-842.

R Core Team. 2013. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

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

Safran M, Dalah I, Alexander J, Rosen N, Iny Stein T, Shmoish M, Nativ N, Bahir I, Doniger T, Krug H et al. 2010. GeneCards Version 3: the human gene integrator. Database (Oxford) 2010:baq020.

Sanford JR, Gray NK, Beckmann K, Cáceres JF. 2004. A novel role for shuttling SR proteins in mRNA translation. Genes Dev 18:755-768.

Savisaar R, Hurst LD. 2016. Purifying Selection on Exonic Splice Enhancers in Intronless Genes. Mol Biol Evol.

Shabalina SA, Spiridonov NA, Kashina A. 2013. Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity. Nucleic Acids Research 41:2073-2094.

Shan J, Munro TP, Barbarese E, Carson JH, Smith R. 2003. A Molecular Mechanism for mRNA Trafficking in Neuronal Dendrites. The Journal of neuroscience 23:8859-8866.

Singh G, Pratt G, Yeo GW, Moore MJ. 2015. The Clothes Make the mRNA: Past and Present Trends in mRNP Fashion. Annu Rev Biochem 84:325-354.

Skrisovska L, Bourgeois CF, Stefl R, Grellscheid SN, Kister L, Wenter P, Elliott DJ, Stevenin J, Allain FH. 2007. The testis-specific human protein RBMY recognizes RNA through a novel mode of interaction. EMBO Rep 8:372-379.

Smith MA, Gesell T, Stadler PF, Mattick JS. 2013. Widespread purifying selection on RNA structure in mammals. Nucleic Acids Research 41:8220-8236.

Smithers B, Oates ME, Gough J. 2015. Splice junctions are constrained by protein disorder. Nucleic Acids Res 43:4814-4822.

Soukarieh O, Gaildrat P, Hamieh M, Drouet A, Baert-Desurmont S, Frebourg T, Tosi M, Martins A. 2016. Exonic Splicing Mutations Are More Prevalent than Currently Estimated and Can Be Predicted by Using In Silico Tools. PLoS Genet 12:e1005756.

Stark A, Brennecke J, Bushati N, Russell RB, Cohen SM. 2005. Animal MicroRNAs Confer Robustness to Gene Expression and Have a Significant Impact on 3′UTR Evolution. Cell 123:1133-1146.

Stergachis AB, Haugen E, Shafer A, Fu W, Vernot B, Reynolds A, Raubitschek A, Ziegler S, LeProust EM, Akey JM et al. 2013. Exonic Transcription Factor Binding Directs Codon Choice and Affects Protein Evolution. Science 342:1367-1372.

Sterne-Weiler T, Howard J, Mort M, Cooper DN, Sanford JR. 2011. Loss of exon identity is a common mechanism of human inherited disease. Genome Res 21:1563-1571.

Sugimoto Y, Konig J, Hussain S, Zupan B, Curk T, Frye M, Ule J. 2012. Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. Genome Biol 13:R67.

Tajnik M, Rogalska ME, Bussani E, Barbon E, Balestra D, Pinotti M, Pagani F. 2016. Molecular Basis and Therapeutic Strategies to Rescue Factor IX Variants That Affect Splicing and Protein Function. PLoS Genet 12:e1006082.

Tournier I, Vezain M, Martins A, Charbonnier F, Baert-Desurmont S, Olschwang S, Wang Q, Buisine MP, Soret J, Tazi J et al. 2008. A large fraction of unclassified variants of the mismatch repair genes MLH1 and MSH2 is associated with splicing defects. Hum Mutat 29:1412-1424.

Tuduri S, Crabbe L, Conti C, Tourriere H, Holtgreve-Grez H, Jauch A, Pantesco V, De Vos J, Thomas A, Theillet C et al. 2009. Topoisomerase I suppresses genomic instability by preventing interference between replication and transcription. Nat Cell Biol 11:1315-1324.

van der Walt S, Colbert SC, Varoquaux G. 2011. The NumPy Array: A Structure for Efficient Numerical Computation. Computing in Science and Engineering:22-30.

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.

Veeramachaneni V, Makałowski W, Galdzicki M, Sood R, Makałowska I. 2004. Mammalian Overlapping Genes: The Comparative Perspective. Genome Res 14:280-286.

Wang Z, Burge CB. 2008. Splicing regulation: From a parts list of regulatory elements to an integrated splicing code. RNA 14:802-813.

Warnecke T, Parmley JL, Hurst LD. 2008. Finding exonic islands in a sea of non-coding sequence: splicing related constraints on protein composition and evolution are common in intron-rich genomes. Genome Biol 9:R29.

Woolfe A, Mullikin JC, Elnitski L. 2010. Genomic features defining exonic variants that modulate splicing. Genome Biol 11.

Wu H, Henras A, Chanfreau G, Feigon J. 2004. Structural basis for recognition of the AGNN tetraloop RNA fold by the double-stranded RNA-binding domain of Rnt1p RNase III. Proc Natl Acad Sci U S A 101:8307-8312.

Wu X, Hurst LD. 2016. Determinants of the Usage of Splice-Associated cis-Motifs Predict the Distribution of Human Pathogenic SNPs. Molecular biology and evolution 33:518-529.

Wu Y, Zhang Y, Zhang J. 2005. Distribution of exonic splicing enhancer elements in human genes. Genomics 86:329-336.

Xing K, He X. 2015. Reassessing the "duon" hypothesis of protein evolution. Mol Biol Evol 32:1056-1062.

Xue Y, Zhou Y, Wu T, Zhu T, Ji X, Kwon YS, Zhang C, Yeo G, Black DL, Sun H et al. 2009. Genome-wide analysis of PTB-RNA interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping. Mol Cell 36:996-1006.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586-1591.

Zhao H, Sun Z, Wang J, Huang H, Kocher JP, Wang L. 2014. CrossMap: a versatile tool for coordinate conversion between genome assemblies. Bioinformatics 30:1006-1007.

Zhou Z, Fu X-D. 2013. Regulation of splicing by SR proteins and SR protein-specific kinases. Chromosoma 122:191-207.