Evolutionary selection on synonymous codons in RNA G-quadruplex structural region

Yuming Xu[1], Ting Qi[1], Zuhong Lu[1,*], Tong Zhou[2,*], Wanjun Gu[1,*]

[1]State Key Laboratory of Bioelectronics, School of Biological Sciences and Medical Engineering, Southeast University, Nanjing, Jiangsu 210096, China
[2]Department of Physiology and Cell Biology, University of Nevada, Reno School of Medicine, Reno, Nevada 89557, USA

[*]Corresponding authors:
Zuhong Lu, State Key Laboratory of Bioelectronics, School of Biological Sciences and Medical Engineering, Southeast University, Sipailou 2, Nanjing, Jiangsu 210096, China. Email: zhlu@seu.edu.cn
Tong Zhou, Department of Physiology and Cell Biology, University of Nevada, Reno School of Medicine, Reno, Nevada 89557, USA; Email: tongz@medicine.nevada.edu
Wanjun Gu, State Key Laboratory of Bioelectronics, School of Biological Sciences and Medical Engineering, Southeast University, Sipailou 2, Nanjing, Jiangsu 210096, China. Email: wanjungu@seu.edu.cn

ABSTRACT

In addition to the amino acid sequence information, synonymous codons can encode multiple regulatory and structural signals in protein coding region. In this study, we investigated how synonymous codons have been adapted to the formation of RNA G-quadruplex (rG4) structure. We found a universal selective pressure acting on synonymous codons to facilitate rG4 formation in five eukaryotic organisms. While *G*-rich codons are preferred in rG4 structural region, *C*-rich codons are selectively unpreferred for rG4 structures. Gene's codon usage bias, nucleotide composition and evolutionary rate can account for the selective variations on synonymous codons among rG4 structures within a species. Moreover, rG4 structures in translational initiation region showed significantly higher selective pressures than those in translational elongation region. These results bring us another dimension of evolutionary selection on synonymous codons for proper RNA structure and function.

INTRODUCTION

mRNAs of protein coding genes in eukaryotic organisms not only carry the genetic information to encode amino acid sequences, but also contain multiple regulatory and structural signals [1]. The pivot provision enabling mRNA to hold these regulatory functions is the redundancy of the genetic code that allows for many "silent" mutations at synonymous codon sites [2]. Although synonymous nucleotide mutations do not change amino acid sequences of the encoded proteins, they can confer dramatic differences to the structure and functions of mRNA itself [1-3]. A wealth of evidence has shown that synonymous codons are selected for optimized RNA stability [4, 5], proper nucleosome positioning [6], efficient mRNA splicing [7, 8], correct microRNA targeting [9], efficient translation initiation [10-12] and elongation [13], and proper protein co-translational folding [14, 15].

The RNA G-quadruplex (rG4) is a non-canonical super-secondary structure around the *G*-rich sites in RNA sequences, which consists of the stacking of G-quartets formed by the G-G Hoogsteen hydrogen bonding [16]. Several experimental studies have shown that rG4 is ubiquitous in eukaryotes, both in untranslated regions (UTRs) and protein coding sequences (CDSs) [17]. RG4 structures have been shown to perform many diverse and vital functions in a wide range of biological processes, such as pre-mRNA splicing [18], alternative polyadenylation [19], mRNA localization [20, 21], microRNA targeting [22], and translational regulation [23-28]. Due to the presence of 2α-hydroxyl property, rG4 structure is more stable than DNA G-quadruplex [29]. In a recent study, Arachchilage *et al.* has demonstrated that the most stable G4s appeared to be significantly under-represented within the CDS by the use of specific synonymous codon combinations [30]. But, several key problems regarding synonymous codon usage and rG4 formation and evolution remain unaddressed.

Here, we looked into the evolutionary choices of synonymous codons around putative rG4 structures (pG4) at the whole transcriptome scale in multiple eukaryotic species. We asked whether there are selective pressures acting on synonymous sites at or around pG4 structures to facilitate the formation of this non-canonical super-secondary structure. If so, how synonymous codons are selected to facilitate rG4 formation? And, what are the factors that may affect this selective pressure? These analyses may help us understand the evolutionary selections acting on synonymous codons in protein coding region.

RESULTS

RG4 structures are selected in protein coding sequences across five eukaryotic species

We calculated the rG4 forming propensity score, $G4Hscore$ [31], along the mRNA sequences using a sliding window scheme. For each pG4 structure in the transcriptome, a 30 nucleotides (nt) window was moved both upstream and downstream from the start position of the pG4 structure with a step of 30 nt, and the $G4Hscore$ of the sequence in each window was calculated for a total of 13 windows. To estimate the background distribution of the formation propensity of rG4 structures, we randomized the mRNA sequences by shuffling synonymous codons for 1,000 times, and calculated the $G4Hscore$ in the corresponding sliding windows. We compared the $G4Hscore$ of the real mRNA sequence in a sliding window with that of 1,000 corresponding sliding windows in the shuffled sequences, and calculated $Z_{G4Hscore}$ to assess the deviation of rG4 formation in the observed mRNA sequence from random expectation (see *Materials and Methods* for details). A positive $Z_{G4Hscore}$ value means synonymous codons are selected to facilitate the formation of rG4 structures, while a negative $Z_{G4Hscore}$ value means a selective pressure that prevents the formation of rG4 structures.

We performed the sliding window analysis in five eukaryotic species, including *Homo sapiens (H. sapiens)*, *Mus musculus (M. musculus)*, *Gallus gallus (G. gallus)*, *Danio rerio (D. rerio)* and *Drosophila melanogaster (D. melanogaster*). We observed a significantly positive $Z_{G4Hscore}$ value in the window of pG4 structures in all five species (Figure 1). This means that synonymous codons are universally selected for the formation of rG4 structures in these five organisms. When sliding windows move to the upstream or downstream of the pG4 structures, the $Z_{G4Hscore}$ values drop quickly in the flank region of pG4 structures, and oscillate around zero for sliding windows far away from the pG4 structures (Figure 1). This suggests that no selective pressures are acting on synonymous codons to facilitate or prevent the formation of rG4 structures when they are located out of the pG4 structures in the genome. When *in vitro* experimentally identified rG4 structures in human *HEK293T* cells and mouse *mESC* cells were used in the analysis, we found a similar pattern of $Z_{G4Hscore}$ changes along the sliding windows (Supplementary Figure S1).

G-rich or C-poor codons are selected for rG4 formation in protein coding region

To investigate how synonymous codons are selected for rG4 formation, we calculated Z_G and Z_C of a window of 30 nt in length for each pG4 structure in protein coding region. When comparing to $Z_{G4Hscore}$ value of the same pG4 structure, we found a

significant positive correlation between Z_G and $Z_{G4Hscore}$ for pG4 structures in all five species (Figure 2). In comparison, a weaker but significant negative correlation between Z_C and $Z_{G4Hscore}$ of pG4 structures was also observed in all five species (Figure 2). These results suggest that synonymous codons with more *G*s and/or less *C*s are generally selected to facilitate the formation of rG4 structures in protein coding region, while the biased usage of *G*-rich codons is more relevant to rG4 formation than that of *C*-poor codons.

Several features of rG4 structure's host gene are associated with rG4 selection in protein coding region

Although the mean value of $Z_{G4Hscore}$ is significantly larger than zero for all exonic pG4 structures in all five species (Figure 1), there are substantial variations among different pG4 structures within a single organism (Figures 1 and 2). To explore the factors that may affect the selective pressures on synonymous codons for rG4 formation, we considered several putative gene-level factors, including codon usage bias, evolutionary rate and nucleotide compositions of the host gene where the pG4 structure is located. We found $Z_{G4Hscore}$ values of pG4 structures in genes with the highest 5% effective number of codons (ENC) are significantly higher than those in genes with the lowest 5% ENC values (Figure 3A). This suggests that pG4 structures in genes with smaller codon usage bias are under stronger selective pressure. For the top 5% genes that have the highest *dN*/*dS* ratio, $Z_{G4Hscore}$ values of pG4 structures in these genes are significantly larger than those of pG4 structures in the bottom 5% genes with the lowest *dN*/*dS* ratio (Figure 3B). When pG4 structures in genes with the highest 5% *G* content are compared to those in genes with the lowest 5% *G* content, we found pG4 structures that located in genes with the highest 5% *G* content had significantly smaller $Z_{G4Hscore}$ values (Figure 3C). Moreover, $Z_{G4Hscore}$ values of pG4 structures in genes with the top 5% *C* content is also statistically smaller than that of pG4 structures in genes with the bottom 5% *C* content (Supplementary Figure S2A). These results suggested that the selective pressures acting on synonymous codons are stronger for pG4 structures in genes with lower *G* nucleotides, less biased usage of synonymous codons and smaller evolutionary rate, and these differences are consistent for pG4 structures in human and mouse (Figure 3).

Selective pressure is different for rG4 structures in different gene regions

Other than their host genes, we also investigated Z_{G4Hscore} differences of rG4 structures in different gene regions. First, we grouped the pG4 structures into two categories, pG4 structures in translation initiation region (within 70nt downstream of

the start codon) and those out of the translation initiation region. We found pG4 structures in the translation initiation region had significantly higher $Z_{\mathrm{G4}H\mathrm{score}}$ values than those in the translation elongation region (Figure 3D). Next, we compared $Z_{\mathrm{G4}H\mathrm{score}}$ values of pG4 structures near exonic splicing sites (within 60 nts of splicing sites) with those far away from the splicing sites. Although $Z_{\mathrm{G4}H\mathrm{score}}$ values for pG4 structures near the splicing sites tend to be smaller than those far away from the splicing sites (Supplementary Figure S2B), the differences are not statistically significant for rG4 structures in both downstream and upstream flank regions of splicing sites. Finally, we parsed pG4 structures in microRNA (miRNA) target sites and compared their $Z_{\mathrm{G4}H\mathrm{score}}$ values with those out of miRNA target region. No significant differences were observed between $Z_{\mathrm{G4}H\mathrm{score}}$ values of pG4 structures in miRNA target region and those out of miRNA target sites (Supplementary Figure S2C). These results suggested synonymous codons in some functional regions, such as translation initiation region, were under different selective pressures for rG4 formation. However, rG4 structures in some regions, including miRNA target region and splicing sites flank region, did not show obviously different evolutionary selections on synonymous codons.

DISCUSSION

In this study, we exploited the usage of synonymous codons in rG4 structural regions and saw obvious selection for the formation of rG4 structures in CDS across all five species (Figure 1). Moreover, *G*-rich and *C*-poor codons are selected in rG4 regions to facilitate the formation of rG4 structures (Figure 2). This is reasonable given the fact that sequences with more *G*s and less *C*s are more prone to form G4 structures, since *C*s may pair with *G*s to competitively form stem-loop structures instead of G4 structures [31]. In a recent study, Archilage *et al.* [30] used rG4 as the model secondary structure to illustrate the impact of codon bias on secondary structure elements within the CDS of mRNAs. They observed that the most stable rG4 structures appeared to be significantly under-represented within the CDS by the use of specific synonymous codon combinations. Their findings strongly lend support to the hypothesis that rG4 structures are selectively avoided in CDS. In comparison to their conclusion, our results suggested synonymous codons, such as G-rich codons, are selected to facilitate rG4 formation when rG4 structures are located in protein coding region, although these ultra-stable RNA structures may be selectively depleted in CDS as they suggested.

We also found several features of rG4 structure's host gene can explain the substantial variations of synonymous selection among rG4 structures within a species (Figure 3). First, genes with higher codon usage bias should have smaller chance to choose specific synonymous codons at specific rG4 structure regions, since synonymous codon usage is more biased and the repertoire of synonymous codons are smaller in these genes. Hence, we observed a reduced selection on synonymous codons in rG4 structures in genes with higher codon usage bias (Figure 3A). Second, genes with higher *dN*/*dS* ratio are more likely to experience positive selection. In comparison to rG4 structures located in conserved genes, rG4 structures in genes with higher evolutionary rate should have higher chance to be positively gained as the evolutionary outcome of recent selection. As expected, we observed a stronger evolutionary selection on synonymous codons for rG4 formation when rG4 structures are located in genes with higher evolutionary rate (Figure 3B). Third, rG4 structures located in genes with higher *G* composition are more likely to use synonymous codons with *G*s inside. Therefore, the selective pressure acting on synonymous codons in these rG4 structures may be smaller given their higher background potential to form rG4 structures (Figure 3C). However, rG4 structures in *C*-rich genes also have weaker evolutionary selection on synonymous codons (Supplementary Figure 2A). This is unexpected since *C*-rich codons are unpreferred in forming rG4 structures (Figure 2). The reason why we observed this difference is unknown, which may be partly explained by the fact that *C* content tends to be equal to *G* content at the genomic level [32].

In addition to above features of their host gene, we also parsed rG4 structures in several specific gene regions to investigate if synonymous codons in these rG4 structures experience significantly different selective pressures, including translational initiation region, exonic splicing sites flank region and miRNA target region (Figure 3D and Supplementary Figure S2). Translation initiation near the start codon has been reported to be the decisive process that determines translation efficiency [10]. Our previous study revealed that reduced secondary structures were selected immediately downstream start codon to facilitate efficient translation [11]. Several rG4 structures were observed to be located near the start codon, which can modulate gene expression by preventing the efficient recognition of *AUG* codon and blocking translational initiation [33]. In our results, we observed that rG4 structures near start codon were under stronger evolutionary selection than other regions along the coding sequence (Figure 3D). This hints the important functions of rG4 structures in translation initiation region in gene expression regulation. In addition, several studies have reported that not only splicing

sites in introns but also those in exons were influenced by the presence of rG4 structures [18, 34-38]. The disruption of rG4 structure was found to inhibit rG4-dependent alternative splicing and cause thousands of alternative splicing events in human cell lines [39]. However, we did not observe a significant different selective pressure when rG4 structures are located in the vicinity of exonic splicing sites (Supplementary Figure S2B). Similarly, we did not observe different selective pressures for rG4 structures in miRNA target region (Supplementary Figure S2C), although rG4 structures could play a role in miRNA-mediated gene regulation by regulating miRNA binding [40, 41]. The reasons why we did not observe significant different selective pressures on synonymous codons in exonic splicing site flank region or miRNA binding region are not known. One possible explanation could be that functional rG4 structures near the exonic splicing sites or miRNA target region are specific to some important regulators, rather than most rG4 structures in these regions.

In conclusion, our results suggested that synonymous codons were selected to facilitate rG4 structure formation and function in protein coding genes. This brought us another dimension of evolutionary selection that acts on synonymous codons in protein coding region.

MATERIALS AND METHODS

Data

We downloaded the nucleotide sequences and exonic structures for all protein coding genes in five eukaryotic species, including *H. sapiens* (GRCh38.p13), *M. musculus* (GRCg6a), *G. gallus* (GRCm38.p6), *D. rerio* (GRCz11) and *D. melanogaster* (BDGP6.28), using *Ensembl BioMarts* (release 97, accessed in July, 2019) [42]. We only considered protein coding genes with a coding sequence more than 150 nts in length. The dN and dS values of all human and mouse orthologous genes were retrieved from *Ensembl BioMarts* as well [43]. Furthermore, we parsed miRNA target sites in protein coding regions of human and mouse genomes from *miRDB* database [44, 45].

To explore the factors that affect the selection for rG4 structure formation, we considered gene codon usage bias, nucleotide composition and its evolutionary rate. We used the effective number of codons (*ENC*) as the measure of codon usage bias, and calculated *ENC* values for each gene as suggested by Wright [46]. A lower *ENC* value indicates stronger codon bias [46]. For each gene, we also calculated nucleotide compositions, including *G* content and *C* content. Moreover, we calculated the ratio of *dN* and *dS* values for each gene, and used the *dN*/*dS* value as the measurement of

evolutionary rate. A *dN*/*dS* value close to zero means the gene is conserved between human and mouse.

rG4 structure in protein coding region

To locate rG4 structures in the protein coding region, we used the *G4Hunter* algorithm [31] to systematically search for potential rG4 forming sites in protein coding sequences in all five species. We ran *G4HunterApps* with a window size at 25 nt and a cutoff at 1.2 [47] (https://github.com/LacroixLaurent/G4HunterMultiFastaSeeker) to identify all potential rG4 structures (pG4) in all protein coding sequences. We chose *G4Hunter* algorithm with these two parameters, since a comprehensive evaluation of computational methods for rG4 prediction has suggested that *G4Hunter* has the best performance in predicting G4 structures [48]. We identified 65,562, 42,153, 38,311, 25,194, and 14,184 rG4 structures in protein coding sequences for *Homo sapiens*, *Mus musculus*, *Gallus gallus*, *Danio rerio*, and *Drosophila melanogaster*, respectively. To validate the observations that we found in predicted rG4 structures, we also downloaded experimentally derived rG4 sites in human *HEK293T* cells and mouse *mESC* cells by high throughput RT-stop techniques from supplemental materials of Guo *et al.* [49].

mRNA Randomization

If the choice of synonymous codons influences the formation of rG4 structures in coding sequences, the *G4Hscore* of mRNA sequences in naturally occurred pG4 region should be statistically different from that of those randomized sequences. Thus, we randomly shuffled synonymous codons in the coding sequence, keeping the amino acid sequence, *GC* composition and codon usage bias the same. For each CDS sequence, the shuffling process was repeated 1,000 times to obtain a set of randomized artificial sequences. The *G4Hscore* of each 30nt window in the native CDS sequence and each permutated sequence were calculated using *G4Hunter* algorithm [31]. The difference of G4 forming potential between the native sequence and the randomized sequences was determined by calculating the *Z*-score of *G4Hscore* for each sliding window by

$$Z_{G4\mathrm{H}score} = \frac{G4Hscore_N - \overline{G4Hscore_P}}{\sqrt{\sum_{i=1}^{n} \frac{\left(G4Hscore_{Pi} - \overline{G4Hscore_P}\right)^2}{n-1}}}$$

Here, $G4Hscore_N$ is the *G4Hscore* for the native sequence in the window, $G4Hscore_{Pi}$ is the *G4Hscore* of the corresponding window of the i^{th} randomized

sequence, and $\overline{G4Hscore_P}$ is the mean of $G4Hscore_{Pi}$ over all randomized sequences. The variable *n* represents the total number of randomized sequences, which is 1,000 here.

Similarly, the difference between the *G* (or *C*) compositions of the native sequence and the randomized sequences was evaluated. The *Z*-score of *G* (or *C*) content (Z_G or Z_C) for each sliding window can be calculated by

$$Z_G = \frac{G_N - \overline{G_P}}{\sqrt{\sum_{i=1}^{n} \frac{\left(G_{Pi} - \overline{G_P}\right)^2}{n-1}}}$$

and

$$Z_C = \frac{C_N - \overline{C_P}}{\sqrt{\sum_{i=1}^{n} \frac{\left(C_{Pi} - \overline{C_P}\right)^2}{n-1}}}$$

The meanings of $G_N\,(C_N)$, $G_{Pi}\,(C_{Pi})$ and $\overline{G_P}\ (\overline{C_P})$ are analogous to those of *G4Hscore*, but refer to *G* or *C* content instead of G4 forming potential.

ACKNOWLEDGEMENTS

This work was funded by grants from National Key R&D Program of China (2018YFC1314900, 2018YFC1314902), National Natural Science Foundation of China (61571109), and the Fundamental Research Funds for the Central Universities (2242017K3DN04).

REFERENCE

1. Shabalina SA, Spiridonov NA, Kashina A. Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity. Nucleic acids research 2013; 41:2073-94.
2. Plotkin JB, Kudla G. Synonymous but not the same: the causes and consequences of codon bias. Nature reviews Genetics 2011; 12:32-42.
3. Hunt RC, Simhadri VL, Iandoli M, Sauna ZE, Kimchi-Sarfaty C. Exposing synonymous mutations. Trends in genetics : TIG 2014; 30:308-21.
4. Chamary J-V, Hurst LD. Biased codon usage near intron-exon junctions: selection on splicing enhancers, splice-site recognition or something else? Trends in Genetics 2005; 21:256-9.
5. Stoletzki N. Conflicting selection pressures on synonymous codon use in yeast suggest selection on mRNA secondary structures. BMC evolutionary biology 2008; 8:224.
6. Warnecke T, Batada NN, Hurst LD. The Impact of the Nucleosome Code on Protein-Coding Sequence Evolution in Yeast. PLoS Genetics 2008; 4.
7. Parmley J, Chamary J, Hurst L. Evidence for Purifying Selection Against Synonymous Mutations

in Mammalian Exonic Splicing Enhancers. Molecular biology and evolution 2006; 23:301-9.

8. Warnecke T, Hurst LD. Evidence for a trade-off between translational efficiency and splicing regulation in determining synonymous codon usage in Drosophila melanogaster. Molecular biology and evolution 2007; 24:2755-62.

9. Gu W, Wang X, Zhai C, Xie X, Zhou T. Selection on synonymous sites for increased accessibility around miRNA binding sites in plants. Molecular biology and evolution 2012; 29:3037-44.

10. Kudla G, Murray AW, Tollervey D, Plotkin JB. Coding-Sequence Determinants of Gene Expression in <em>Escherichia coli</em&gt. Science 2009; 324:255.

11. Gu W, Zhou T, Wilke CO. A Universal Trend of Reduced mRNA Stability near the Translation-Initiation Site in Prokaryotes and Eukaryotes. PLoS computational biology 2010; 6:e1000664.

12. Tuller T, Waldman YY, Kupiec M, Ruppin E. Translation efficiency is determined by both codon bias and folding energy. Proc Natl Acad Sci U S A 2010; 107:3645-50.

13. Gingold H, Pilpel Y. Determinants of translation efficiency and accuracy. Molecular Systems Biology 2011; 7.

14. Thanaraj T, Argos P. Ribosome-mediated translational pause and protein domain organization. Protein science : a publication of the Protein Society 1996; 5:1594-612.

15. Komar A, Lesnik T, Reiss C. Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. FEBS letters 2000; 462:387-91.

16. Kharel P, Balaratnam S, Beals N, Basu S. The role of RNA G-quadruplexes in human diseases and therapeutic strategies. Wiley Interdisciplinary Reviews: RNA, 2020:1-20.

17. Kwok C, Marsico G, Sahakyan AB, Chambers VS, Balasubramanian S. rG4-seq reveals widespread formation of G-quadruplex structures in the human transcriptome. Nature methods 2016; 13:nmeth.3965.

18. Huang H, Zhang J, Harvey SE, Hu X, Cheng C. RNA G-quadruplex secondary structure promotes alternative splicing via the RNA-binding protein hnRNPF. Genes and Development 2017; 31:2296-309.

19. Beaudoin JD, Perreault JP. Exploring mRNA 3′-UTR G-quadruplexes: Evidence of roles in both alternative polyadenylation and mRNA shortening. Nucleic acids research 2013; 41:5898-911.

20. Subramanian M, Rage F, Tabet R, Flatter E, Mandel J-L, Moine H. G-quadruplex RNA structure as a signal for neurite mRNA targeting. EMBO reports 2011; 12:697-704.

21. Kanai Y, Dohmae N, Hirokawa N. Kinesin Transports RNA: Isolation and Characterization of an RNA-Transporting Granule. Neuron 2004; 43:513-25.

22. Stefanovic S, Bassell GJ, Mihailescu MR. G quadruplex RNA structures in PSD-95 mRNA: Potential regulators of miR-125a seed binding site accessibility. Rna 2015; 21:48-60.

23. Beaudoin J-DD, Perreault J-PP. 5'-UTR G-quadruplex structures acting as translational repressors. Nucleic acids research 2010; 38:7022-36.

24. Bugaut A, Balasubramanian S. 5'-UTR RNA G-quadruplexes: translation regulation and targeting. Nucleic acids research 2012; 40:4727-41.

25. Kamura T, Katsuda Y, Kitamura Y, Ihara T. G-quadruplexes in mRNA: A key structure for biological function. Biochemical and Biophysical Research Communications 2020; 526:261-6.

26. Kumari S, Bugaut A, Huppert JL, Balasubramanian S. An RNA G-quadruplex in the 5′ UTR of the NRAS proto-oncogene modulates translation. Nature Chemical Biology 2007; 3:218-21.

27. Murat P, Marsico G, Herdy B, Ghanbarian AT, Portella G, Balasubramanian S. RNA

G-quadruplexes at upstream open reading frames cause DHX36- and DHX9-dependent translation of human mRNAs. Genome biology 2018; 19:229.

28. Simone R, Fratta P, Neidle S, Parkinson GN, Isaacs AM. G-quadruplexes: Emerging roles in neurodegenerative diseases and the non-coding transcriptome. FEBS letters, 2015:1653-68.

29. Fay MM, Lyons SM, Ivanov P. RNA G-quadruplexes in biology: principles and molecular mechanisms. Journal of Molecular Biology: Elsevier Ltd, 2017:2127-47.

30. Mirihana Arachchilage G, Hetti Arachchilage M, Venkataraman A, Piontkivska H, Basu S. Stable G-quadruplex enabling sequences are selected against by the context-dependent codon bias. Gene 2019; 696:149-61.

31. Bedrat A, Lacroix L, Mergny J-L. Re-evaluation of G-quadruplex propensity with G4Hunter. Nucleic acids research 2016; 44:1746-59.

32. Prabhu VV. Symmetry observations in long nucleotide sequences. Nucleic acids research 1993; 21:2797-800.

33. Wieland M, Hartig JS. RNA quadruplex-based modulation of gene expression. Chem Biol 2007; 14:757-63.

34. Fisette J-F, Montagna D, Mihailescu R, Wolfe M. A G-Rich element forms a G-quadruplex and regulates BACE1 mRNA alternative splicing. Journal of neurochemistry 2012; 121:763-73.

35. Kikin O, Zappala Z, D'Antonio L, Bagga P. GRSDB2 and GRS_UTRdb: databases of quadruplex forming G-rich sequences in pre-mRNAs and mRNAs. Nucleic acids research 2008; 36:D141-8.

36. Marcel V, Tran P, Sagne C, Martel-Planche G, Vaslin L, Teulade-Fichou M-P, et al. G-quadruplex structures in TP53 intron 3: Role in alternative splicing and in production of p53 mRNA isoforms. Carcinogenesis 2010; 32:271-8.

37. Gomez D, Lemarteleur T, Lacroix L, Patrick M, Mergny J-L, Riou J-F. Telomerase downregulation induced by the G-quadruplex ligand 12459 in A549 cells is mediated by hTERT RNA alternative splicing. Nucleic acids research 2004; 32:371-9.

38. Didiot M-C, Tian Z, Schaeffer C, Subramanian M, Mandel J-L, Moine H. The G-quartet containing FMRP binding site in FMR1 mRNA is a potent exonic splicing enhancer. Nucleic acids research 2008; 36:4902-12.

39. Zhang J, Harvey SE, Cheng C. A high-throughput screen identifies small molecule modulators of alternative splicing by targeting RNA G-quadruplexes. Nucleic acids research 2019; 47:3667-79.

40. Kharel P, Becker G, Tsvetkov V, Ivanov P. Properties and biological impact of RNA G-quadruplexes: from order to turmoil and back. Nucleic acids research 2020; 48:12534-55.

41. Rouleau S, Glouzon JPS, Brumwell A, Bisaillon M, Perreault JP. 3′ UTR G-quadruplexes regulate miRNA binding. Rna: Cold Spring Harbor Laboratory Press, 2017:1172-9.

42. Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, et al. Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database 2011; 2011:bar030.

43. Herrero J, Muffato M, Beal K, Fitzgerald S, Gordon L, Pignatelli M, et al. Ensembl comparative genomics resources. Database 2016; 2016:bav096.

44. Chen Y, Wang X. miRDB: an online database for prediction of functional microRNA targets. Nucleic acids research 2020; 48:D127-D31.

45. Liu W, Wang X. Prediction of functional microRNA targets by integrative modeling of microRNA binding and target expression data. Genome biology 2019; 20:18.

46. Wright F. The ‘effective number of codons’ used in a gene. Gene 1990; 87:23-9.

47. Lacroix L. G4HunterApps. Bioinformatics 2018; 35:2311-2.

48. Lombardi EP, Londono-Vallejo A. A guide to computational methods for G-quadruplex prediction. Nucleic acids research 2020; 48:1603.

49. Guo JU, Bartel DP. RNA G-quadruplexes are globally unfolded in eukaryotic cells and depleted in bacteria. Science 2016; 353.

FIGURE LEGENDS

Figure 1. Standard error and mean value of $Z_{G4Hscore}$ of 13 sliding windows in protein coding sequences in five eukaryotic species. The sliding windows are centered at pG4 structures in protein coding sequences, and moved both upwards and downwards for six windows with an offset at 30nts in length.

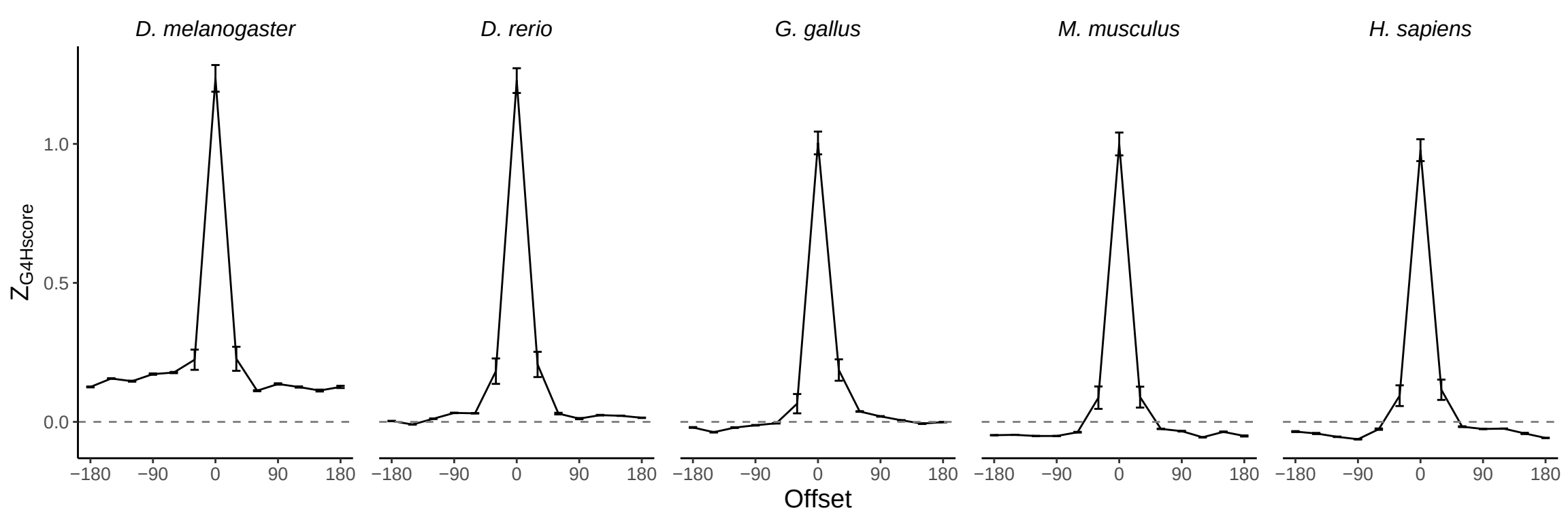
Figure 2. The correlations between $Z_{G4Hscore}$ value and Z_G **(A)** or Z_C **(B)** value for each pG4 structure in protein coding region in all five eukaryotic species.

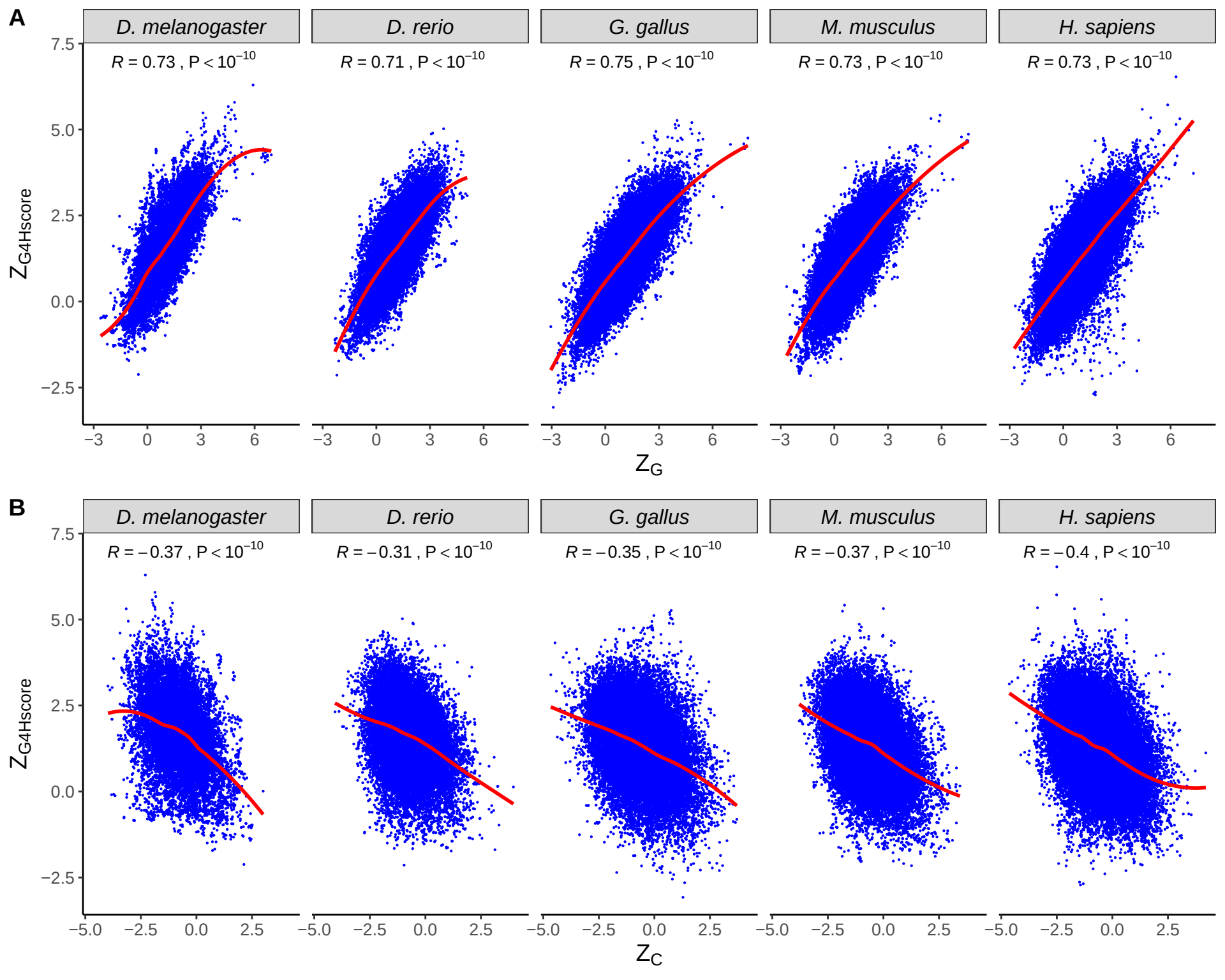
Figure 3. Factors that may explain $Z_{G4Hscore}$ variations among rG4 structures in protein coding region of human and mouse, including *ENC* value **(A)**, *dN*/*dS* value **(B)** and *G* content **(C)** of the host gene of rG4 structure. Moreover, rG4 structures in translation initiation region have significantly higher $Z_{G4Hscore}$ values than those in translation elongation region **(D).**

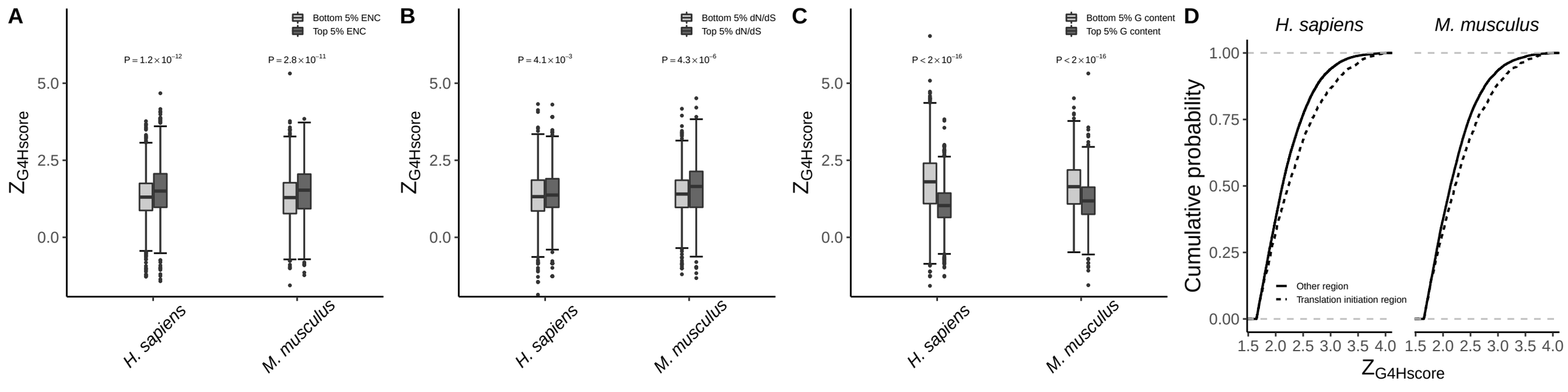
SUPPLEMENTARY FILES

Supplementary Figure 1. Standard error and mean of $Z_{G4Hscore}$ of 13 sliding windows in protein coding sequences in human and mouse. The sliding windows are centered at *in vitro* experimentally validated rG4 structures in protein coding sequences, and moved both upwards and downwards for six windows with an offset at 30nts in length.

Supplementary Figure 2. **(A)** *C* content of the host gene of rG4 structure may not explain $Z_{G4Hscore}$ variations among rG4 structures in protein coding region of human and mouse. In addition, when comparing to rG4 structures in other protein coding regions, rG4 structures in the flank region of exonic splicing sites **(B)** and miRNA target region **(C)** did not show significantly different $Z_{G4Hscore}$ values**.**


D. melanogaster
D. rerio
G. gallus
M. musculus
H. sapiens
$Z_{G4Hscore}$
1.0
0.5
0.0
−180
−90
0
90
180
Offset


A
D. melanogaster
D. rerio
G. gallus
M. musculus
H. sapiens
R = 0.73 , P < 10−10
R = 0.71 , P < 10−10
R = 0.75 , P < 10−10
R = 0.73 , P < 10−10
R = 0.73 , P < 10−10
ZG4Hscore
ZG
B
D. melanogaster
D. rerio
G. gallus
M. musculus
H. sapiens
R = −0.37 , P < 10−10
R = −0.31 , P < 10−10
R = −0.35 , P < 10−10
R = −0.37 , P < 10−10
R = −0.4 , P < 10−10
ZG4Hscore
ZC


A
Bottom 5% ENC
Top 5% ENC
P = 1.2×10−12
P = 2.8×10−11
ZG4Hscore
5.0
2.5
0.0
H. sapiens
M. musculus
B
Bottom 5% dN/dS
Top 5% dN/dS
P = 4.1×10−3
P = 4.3×10−6
C
Bottom 5% G content
Top 5% G content
P < 2×10−16
P < 2×10−16
D
H. sapiens
M. musculus
Cumulative probability
1.00
0.75
0.50
0.25
0.00
Other region
Translation initiation region
1.5 2.0 2.5 3.0 3.5 4.0
ZG4Hscore