4 Chapter IV: Analysis of non-B DNA motifs across splice sites

Collaboration note

The work described in this chapter is currently published as pre-print in biorxiv (Georgakopoulos-Soares et al.) and under review in Nature Communications. Both Ilias Georgakopoulos-Soares and I conceived the project, and contributed equally to the manuscript and bioinformatic analyses. While Ilias Georgakopoulos-Soares quantified non-B DNA motifs across the genome, I processed splice sites and took the lead compiling an R notebook that generates all of the figures presented here. All of our work was done under the supervision of Martin Hemberg and Eric Miska.

4.1 Introduction

Nucleic acid oligomers can adopt different conformations. In the case of DNA the most frequent structure formed *in vivo* corresponds to B-DNA, a right handed double helix, which is considered the canonical DNA structure. However, different sequence contexts, environments and biological processes, such as replication and transcription, can favour the formation of non-canonical DNA structures collectively known as non-B DNA. More than 20 non-canonical secondary structures have been previously reported for DNA, including G-quadruplexes (G4s), hairpins, cruciforms and triplexes (Ghosh and Bansal, 2003).

Sequences that predispose DNA to non-canonical conformations are known as non-B DNA motifs and they have been associated to different sources of genome instability, such as translocations and double strand breaks (Bacolla et al., 2016; Georgakopoulos-Soares et al., 2018). However, some of them can also have regulatory roles of gene expression. In particular, G4s have been shown to be enriched in promoters and nucleosome depleted regions, and some of them have

been found to have important gene regulation roles (Hänsel-Hertsch et al., 2016; Huppert and Balasubramanian, 2007). For example, a G4 in the promoter of the oncogene *MYC* acts as a repressor (Hurley et al., 2006; Siddiqui-Jain et al., 2002; Yang and Hurley, 2006). Similarly, a G4 in the promoter of the proto-oncogene *KRAS* has a negative effect on expression levels (Cogoi and Xodo, 2006).

Since many non-B DNA motifs can also lead to similar secondary structures at the RNA level, their formation has the potential to affect mRNA processing (Bevilacqua et al., 2016a; Kwok and Merrick, 2017; Uzilov and Underwood, 2016). Since the presence of an extra 2′-hydroxyl group on RNA molecules promotes additional intramolecular interactions within RNA G4s, G4s are more stable in RNA than DNA molecules (Fay et al., 2017; Zhang et al., 2010). However, the impact of non-canonical DNA and RNA structures over alternative splicing remains only partially understood (Buratti and Baralle, 2004; Warf and Berglund, 2010) and although a role of G4s in splicing has been suggested (Gomez et al., 2004; Hastings and Krainer, 2001; Huang et al., 2017; Marcel et al., 2011; Tsai et al., 2014; Weldon et al., 2018; Zhang et al., 2019a), the extent of G4 impact on alternative splicing remains to be explored. In this chapter I describe a systematic characterization of non-B DNA motifs across splice sites and an exploration of their implications for alternative splicing modulation.

4.2 Results

4.2.1 Genome wide analysis of non-B DNA motifs across splice sites

To investigate the contribution of non-canonical secondary structures to splice site definition, we systematically explored the distribution of seven known non-B DNA motifs. Since the secondary structures can form both at the DNA and RNA level (Bevilacqua et al., 2016b; Biffi et al., 2013; Strobel et al., 2018) and it is plausible that DNA structures could have an impact, both strands were considered for this initial analysis. In order to characterize the distribution of non-B DNA motifs across

splice sites, we calculated the enrichment of non-B DNA motif occurrences across splice site flanking regions. The enrichment profiles varied substantially across the different non-B DNA motif categories (Fig 4.1), with the highest enrichment found for G4s, both at the 3′ss (2.44-fold) and the 5′ss (4.06-fold). High enrichment of short tandem repeats was also observed, but that was expected since a subset of them overlap with intronic polypyrimidine tracts which are known to be part of the core splicing signal (Coolidge et al., 1997; Dominski and Kole, 1991). By contrast, the enrichment patterns for G4s or H-DNA motifs cannot be explained by the distribution of known splicing signals.

Alternative splice sites are often associated with weak splice sites, which allow them to be modulated by cis-regulatory elemens (Ast, 2004). Thus, if non-B DNA structures function as splicing cis-regulatory elements, their enrichments may vary across exons flanked by weak and strong splice sites. To investigate the association of non-B DNA motifs and splice site strength, we measured the splice strength using publicly available position weight matrices (Sheth et al., 2006), and divided the splice sites into quartiles based on how well they match the position weight matrix. We found that several non-B DNA motifs are not evenly enriched across splice site quartiles (Fig 4.2). While some non-B DNA motifs had higher enrichment at strong splice sites (splice site quartile 4, Q4), G4s were more enriched at weak splice sites, suggesting that they may be associated with alternative splicing events.

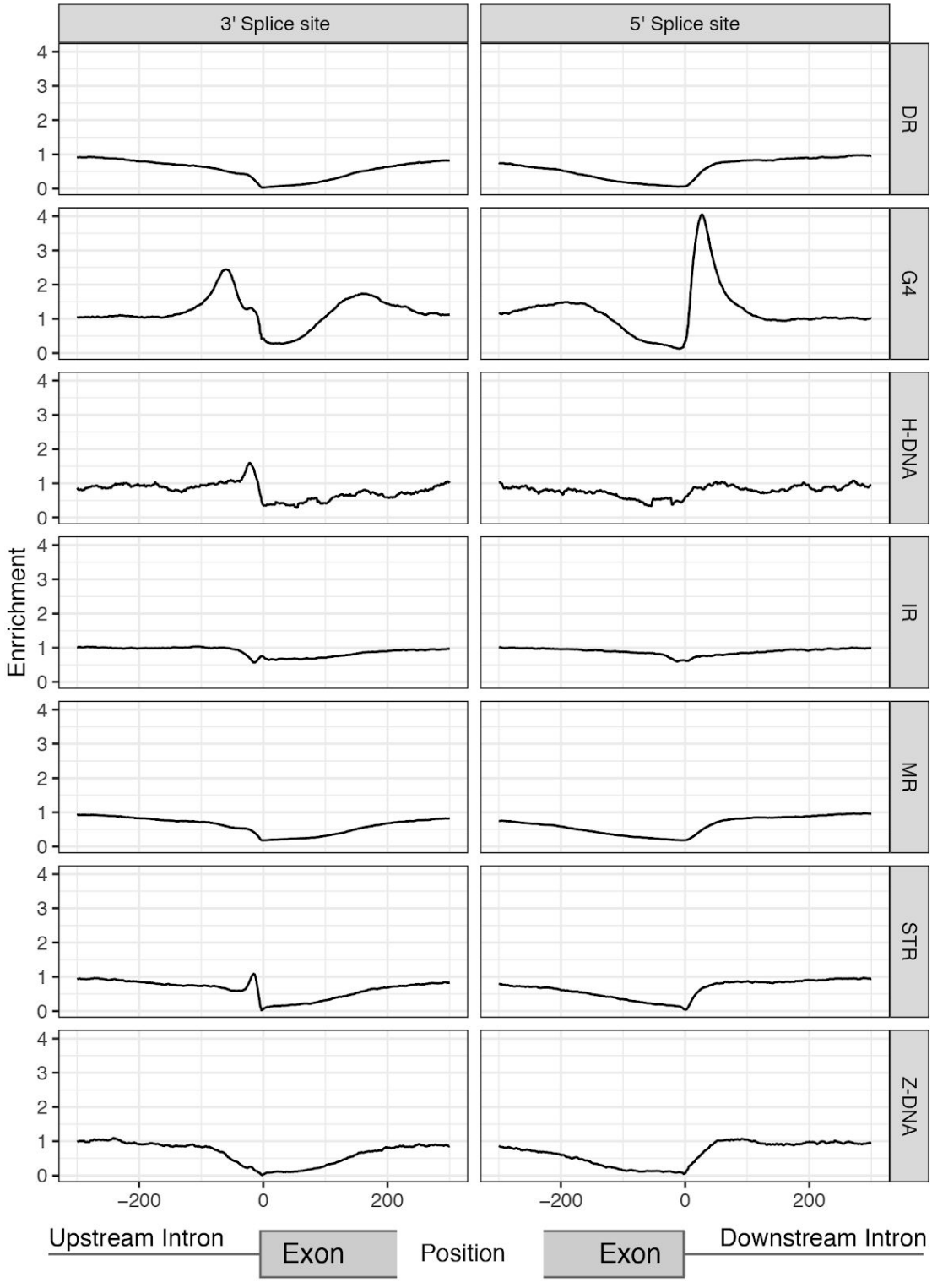

Figure 4.1: Landscape of non-B DNA motifs across human splice sites. Distribution of non-B DNA motifs relative to splice sites. Seven non-B DNA motifs are shown, namely direct repeats (DRs), G-quadruplexes (G4s), H-DNA, inverted repeats (IRs), mirror repeats (MRs), short tandem repeats (STRs) and Z-DNA. Enrichment was calculated as the occurrences of a non-B DNA motif at a given position over the median number of occurrences of that motif across a 1kB window each side from the splice site.

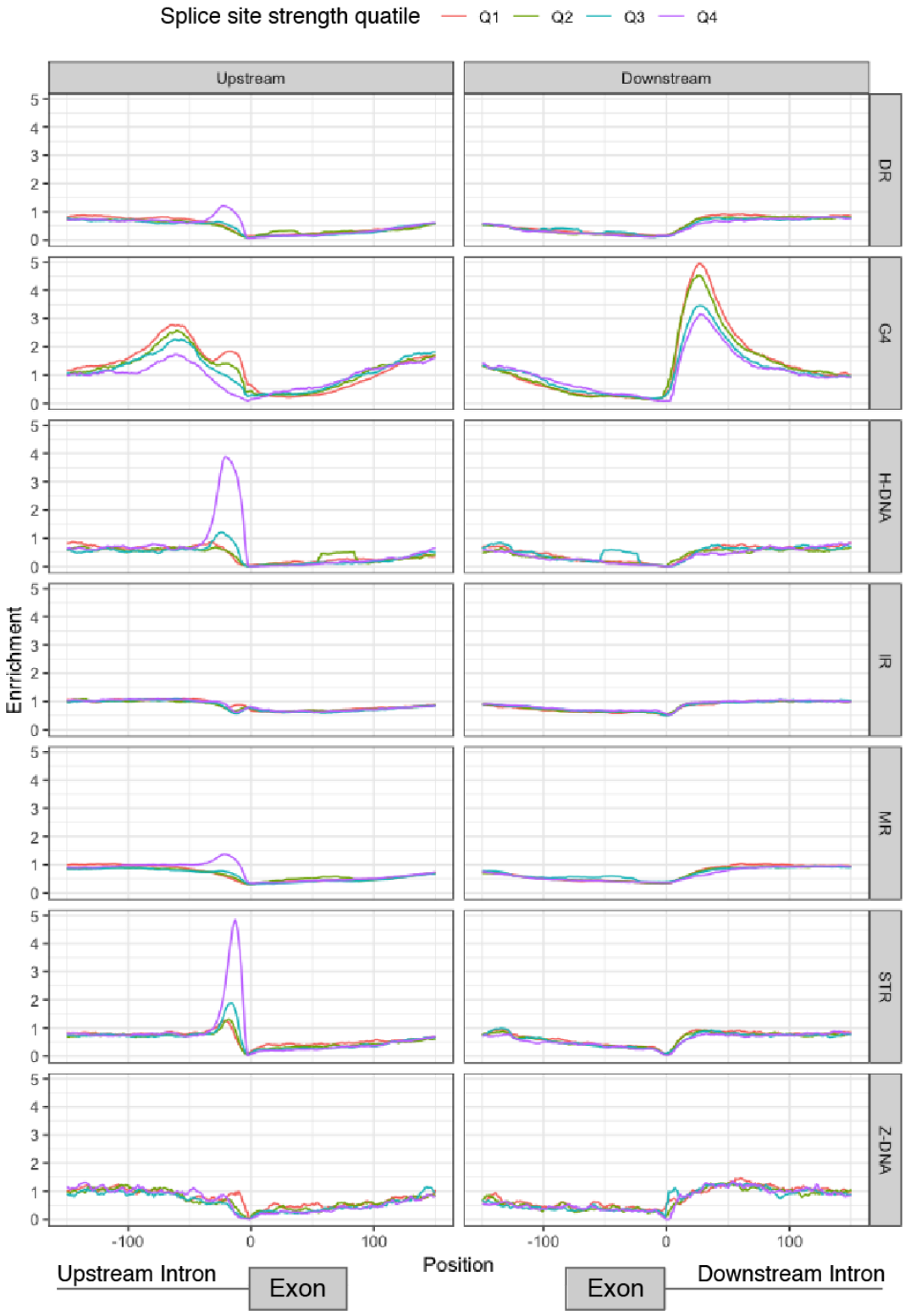

Figure 4.2: Non-B DNA motif enrichment varies with splice strength. Enrichment of non-B DNA motifs across splice site strength quartiles. Seven non-B DNA motifs are shown, namely direct repeats (DRs), G-quadruplexes (G4s), H-DNA, inverted repeats (IRs), mirror repeats (MRs), short tandem repeats (STRs) and Z-DNA.

4.2.2 G4 enrichment analyses of splice sites

Since G4s had the strongest enrichment and they were particularly enriched at weak splice sites (Fig 4.1, 4.2), we investigated the distribution of G4s across splice sites in further detail. To control for the effect of the nucleotide composition of splice sites in the distribution of the GC-rich G4s, we shuffled the 100 nt window each side of the splice site while controlling for dinucleotide content. Comparing the observed frequency to the median from 1,000 permutations we observed a corrected 2.53-fold and 2.73-fold enrichment for the frequency of G4s at the 3′ss and 5′ss, respectively (p-value<0.001 in both 3′ss and 5′ss), indicating that the G4 patterns are not driven by the sequence composition of splice sites. Since G4 motif enrichment was highest across intronic regions that are proximal to splice junctions, we count the number of splice junctions that have at least a G4 motif in close proximity. Within 100 nt of each splice junction we identified 19,987 and 20,088 G4s at the 3′ss and 5′ss, respectively. In total, 31% of human genes contain a G4 motif near at least one splice site within a distance of 100 bp. G4 motifs were found within 100 nt for 8.79% and 8.83% of the 3′ss and 5′ss, respectively. The reported G4 motif frequencies are likely a conservative estimate since we do not take into account intermolecular G4s or G4s that do not adhere to the consensus motif (G□3N1-7G□3N1-7G□3N1-7G□3), (Huppert and Balasubramanian, 2007; Kikin et al., 2006; Varizhuk et al., 2017).

Since G4 formation and template DNA promote polymerase stalling, polymerase stop assays have been implemented to detect G4 formation *in vitro* (Weitzmann et al., 1996). Moreover, since polymerase stalling was found to affect base calling, Chambers and collaborators were able to develop a genome-wide DNA G4 formation assay based on high-throughput DNA sequencing, which they called G4-seq (Chambers et al., 2015). During this assay G4 are stabilised *in vitro* using K+ or pyridostatin (PDS), and G4 formation is inferred by the detection of mismatches induced by base calling errors. Improved versions of the G4-seq protocol have been used to generate maps of G4 formatting sequences across different species, demonstrating a strong enrichment at gene promoter regions and 5′ UTR for human, mouse and *Trypanosoma* (Marsico et al., 2019b). Thus, we analysed these publicly

available G4-seq data to corroborate our findings regarding G4 motif enrichments around splice sites.

We first measured the distribution of G4s relative to the splice sites for HEK-293T cells (a human cell line) in Pyridostatin (PDS) and K^+ treatments from (Marsico et al., 2019b). In both conditions, we observed an enrichment of G4-seq peaks relative to the 3′ss and 5′ss, but with a more pronounced G4 enrichment in PDS treatment compared to K^+ treatment (Fig 4.3a). The majority of G4 positions derived from G4-seq peaks in K^+ and PDS treatments did not overlap consensus G4 motifs (Fig 4.3b), which could be explained by the wider range of G4 structures that can be detected though G4-seq, such as noncanonical long loop and bulged structures that cannot be detected with our conservetive G4 motif definition (Chambers et al., 2015). From all G4 motifs that we predicted *in silico*, at least 66.31% and 66.27% were detected at G4-seq experiments (under K^+ and PDS condition) for the 3′ and 5′ splice sites respectively. We also found 20.88% and 21.05% of overlapping peaks between G4-seq experiments for 3′ and 5′ splice sites respectively.

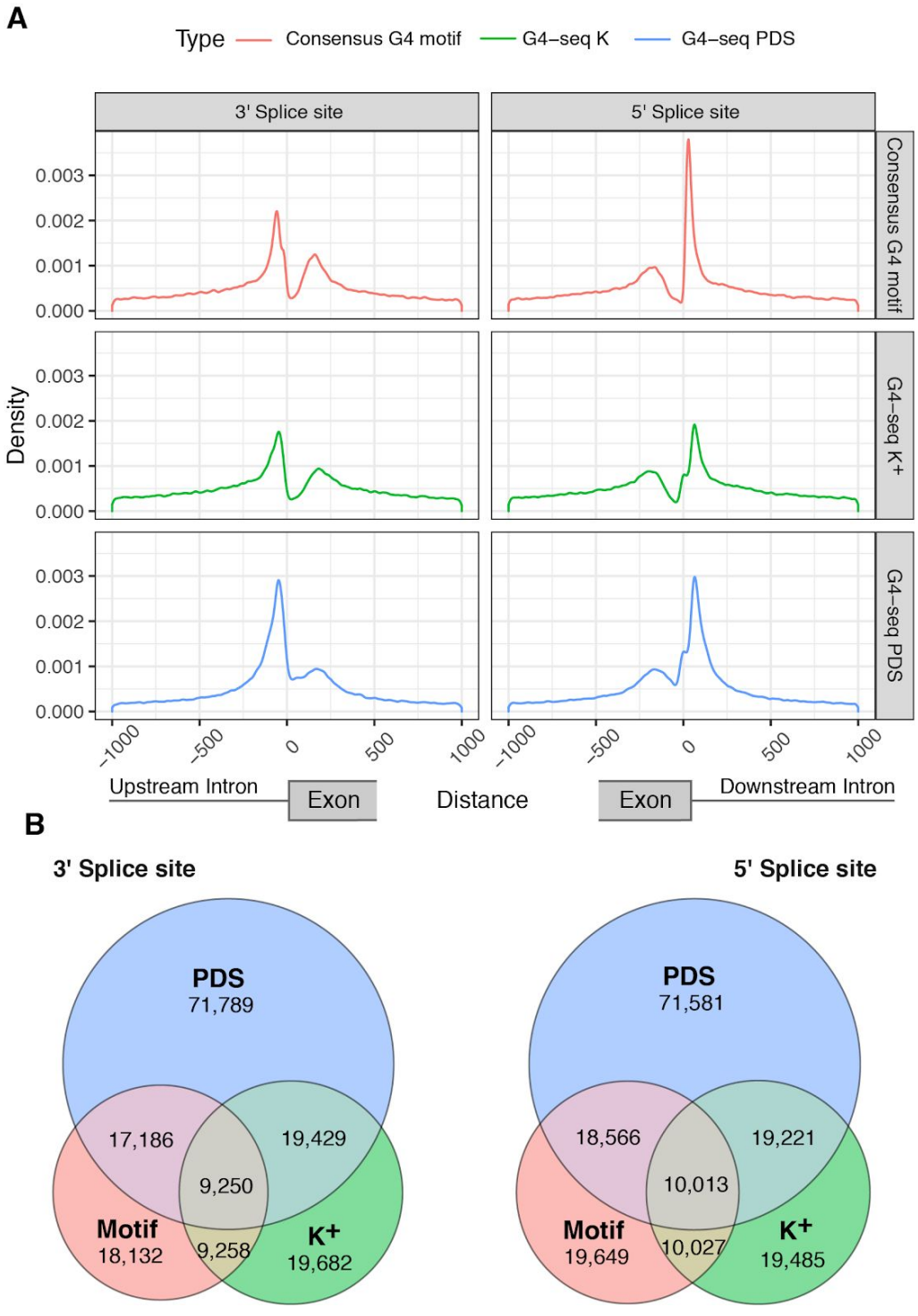

Figure 4.3: Analysis of high-resolution G4-seq data validates in-silico enrichment of G4 motifs A. Distance between nearest G4 motif / G4-seq peak and a splice site separately for 3′ / 5′ splice sites. **B.** Venn diagrams for the occurrences of G4s within 100 nt of the 3′ss (upstream) and 5′ss (downstream) using the consensus G4 motif, the K^+ treatment G4-seq derived G4 peaks and the PDS treatment G4-seq derived G4 peaks (Marsico et al., 2019b) and reporting the overlapping G4s between them.

4.2.2.1 G4s are enriched at weak splice sites

Weak splice sites are highly involved in alternative splicing and often contain additional regulatory elements (Erkelenz et al., 2018; Parada et al., 2014; Sibley et al., 2016). To explore the distribution of G4s across weak and strong splice sites, we calculated a splicing strength score for all internal exons based on splice site position weight matrices (Parada et al., 2014; Sheth et al., 2006). We grouped splice sites into four quantiles based on the splicing strength scores, and explored the enrichment levels of G4-seq peaks for each quantile separately. We found an inverse relationship between the calculated splicing strength score and G4 enrichment, with the weakest splice sites having the highest enrichment of G4s both at the 3′ss and the 5′ss with 2.77-fold and 4.95-fold enrichment, respectively (Fig 4.4a-b). For both mouse and human, the splicing strength scores for splice junctions with a G4 are significantly lower than for splice junctions without a G4 (Mann-Whitney U, p-value<0.001).

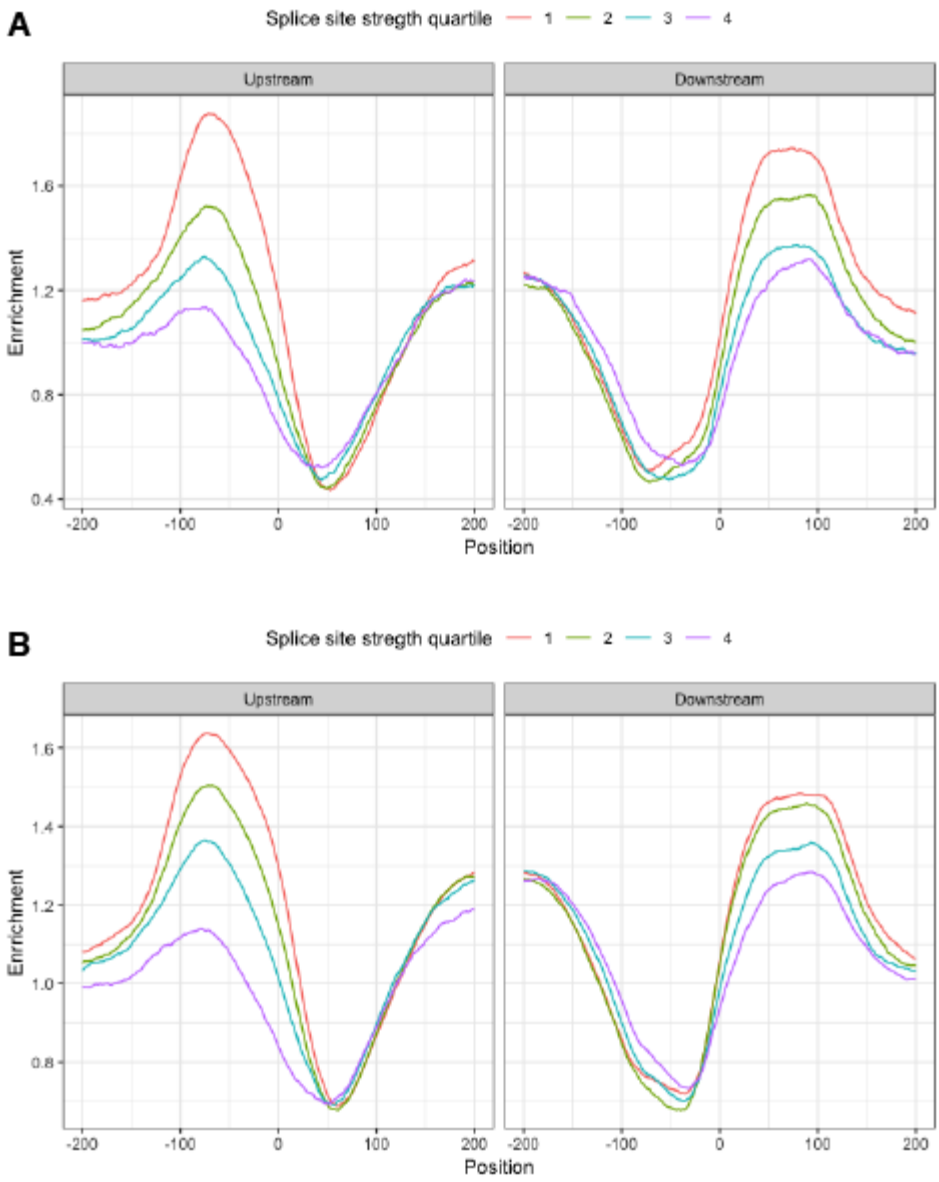

Figure 4.4: Splice site strength and distribution of G4 motifs at splicing sites. G4s display a stronger enrichment at weaker splice sites. A. Distribution of G4 peaks derived from G4-seq with K^+ treatment from (Marsico et al., 2019b) at splicing sites at 3′ss (upstream) and 5′ss (downstream) and the association with splicing strength. B. Distribution of G4 peaks derived from G4-seq with PDS treatment from (Marsico et al., 2019b) at 3′ss (upstream) and 5′ss (downstream) and the association with splicing strength.

4.2.2.2 G4s are preferentially found on the non-template strand

Since G4s are strand specific, we oriented each instance relative to the direction of transcription. Thus, we considered G4s found at the template (non-coding) and non-template (coding) strands separately and found them statistically enriched on the non-template strand (Binomial tests, p-value<0.001 at 3′ss and 5′ss). Moreover, G4s were enriched at both strands relative to flanking sequences (Fig 4.5). At 3′ss the enrichment was 3.01-fold and 2.78-fold enrichment scores at the non-template and template strands, respectively. At the 5′ss the difference between the strands was larger with 5.56-fold and 2.38-fold at the non-template and template strands, respectively. Therefore, there was an asymmetric enrichment between the template and non-template strands at the 5′ss, but only a weak asymmetry at the 3′ss.

We also investigated if there was a strand asymmetry when considering the splicing strength scores. Indeed, we found a bias in the splicing strength scores dependent on the strand orientation of G4s (Mann-Whitney U, 3′ss p-value<0.05, 5′ss p-value<0.001). At the 3′ss the enrichment for splice junctions with the weakest splicing strength scores at the template and non-template strand was 3.90-fold and 3.66-fold, respectively. By contrast, we observed a 6.76-fold enrichment for G4s at the 5′ss at the non-template strand, but only a 3.66-fold enrichment on the template strand at the splice junctions with the weakest splicing strength scores (Fig 4.5).

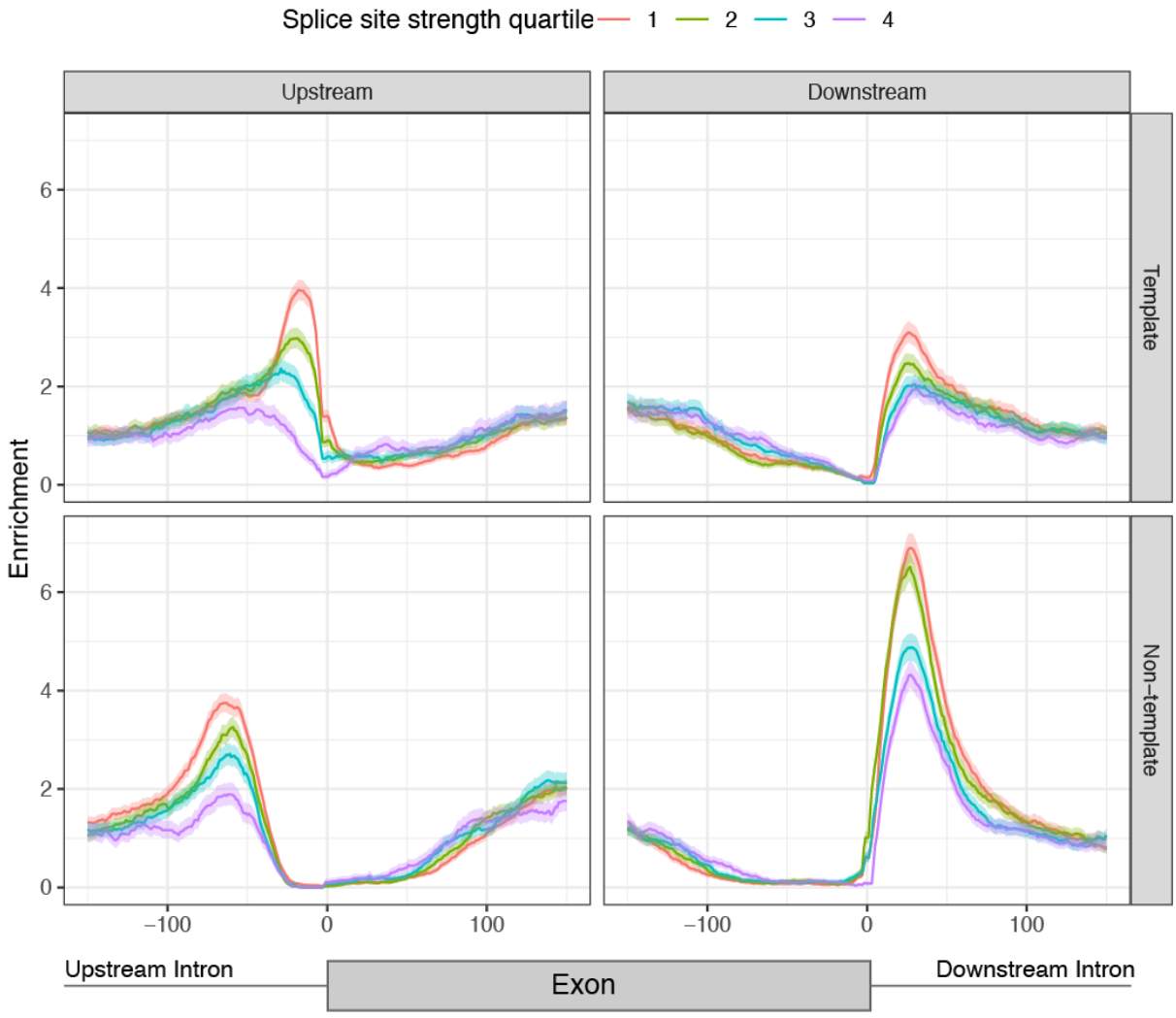

Figure 4.5: Characterisation of G4 motifs across splicing junctions. A. G4 enrichment for template and non-template strands and stratified by the splicing strength scores of the adjacent splice site. The splicing strength scores for splice junctions with a G4 are significantly lower than for splice junctions without a G4 (Mann-Whitney U, p-value<0.001). The splicing strength score bias was found to be dependent on the strand orientation of G4s for splice sites with G4s within 100 nt away (Mann-Whitney U, 3′ss p-value<0.05, 5′ss p-value<0.001).

4.2.2.2.1 Replicated effects are found in independent G4-seq experiments

In order to corroborate the observed G4 motif and G4-seq peak enrichment (Marsico et al., 2019b) patterns, we used additional G4-seq data produced independently. Chambers and collaborators performed a G4-seq experiment in primary human B lymphocytes (NA18507) under Na^+-K^+ or Na^+-PDS conditions (Chambers et al., 2015), both of which promote G4 formation. Even though both Chambers et. al and Marsico et al. data were performed using different cell lines and different ionic solution as G4 stabilized treatments, both G4-seq data show similar enrichment patterns across splice site strength quartiles, displaying the same inverse relationship between splice strength quartile and enrichment (Fig 4.6) (Mann-Whitney U, p-values<0.001).

For both the PDS and K^+ treatments we find that a substantial fraction of the genome is affected, with 31.72% and 10.25% of splice junctions having a G4 within 100 nt intronic window. In addition, 67% and 35% of human genes contain a G4-seq peak from PDS and K^+ treatments within 100 nt of a splice junction, supporting our earlier observations using the consensus G4 motif. As a result of these findings, we conclude that G4s are a pervasive feature near splicing junctions.

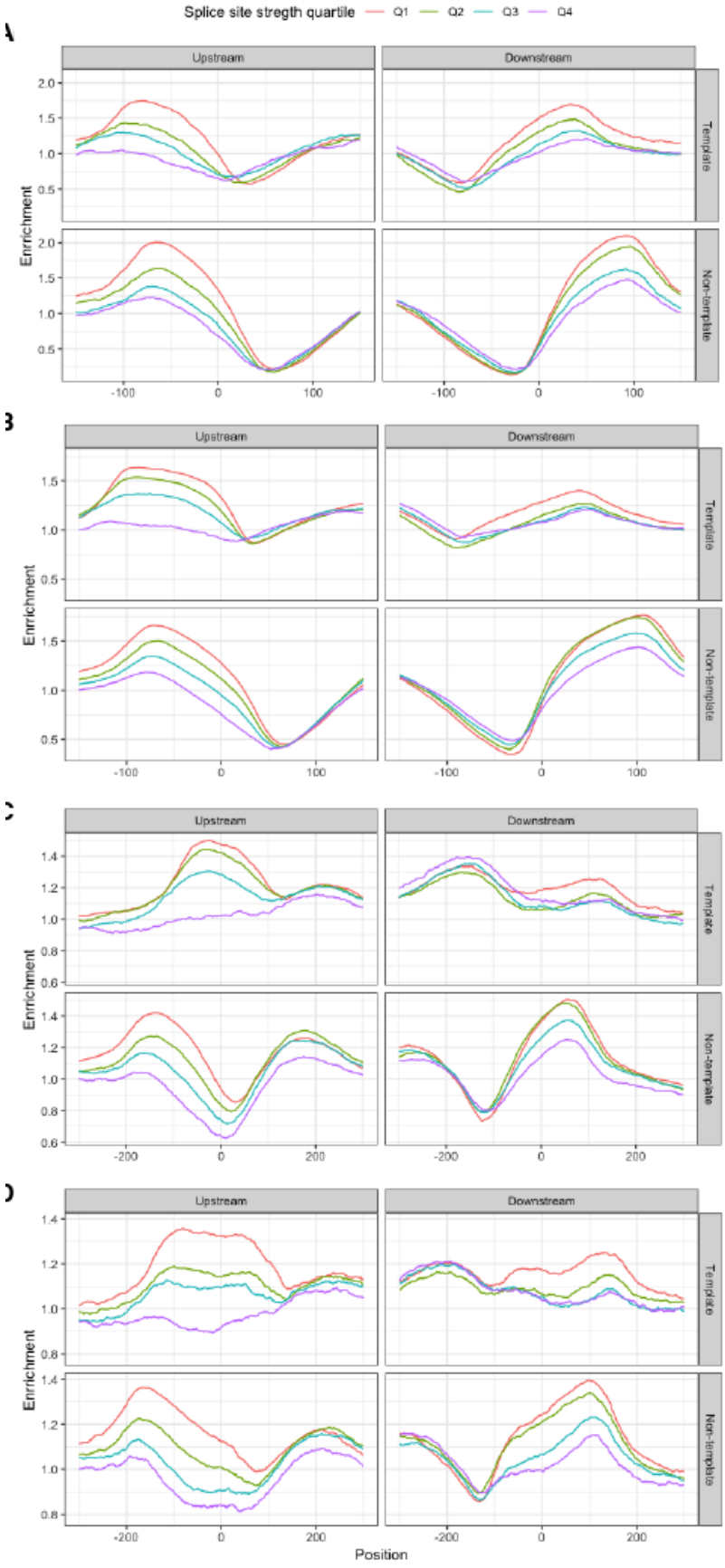

Figure 4.6: G4 enrichment patterns are consistent across different G4-seq experiments. A. Distribution of G4 peaks derived from G4-seq in presence of PDS at template and non-template strands and the association with splicing strength. **B.** Distribution of G4 peaks derived from G4-seq in presence of K^+ at template and non-template strands and the association with splicing strength. **C.** Distribution of G4 peaks derived from G4-seq in presence of Na^+-PDS at template and non-template strands and the association with splicing strength. **D.** Distribution of G4 peaks derived from G4-seq in presence of Na^+-K^+ at template and non-template strands and the association with splicing strength. **A-B.** Are based on recent G4-seq data from (Marsico et al., 2019b) with higher resolution than **C-D** (Chambers et al., 2015). But the same trend is shown, validating the observation across two independent sets of experiments.

4.2.2.2.2 Template and non-template G4 enrichment patterns across gene body

Since the enrichment of G4s around promoters have been reported to be biased towards the non-template DNA strand (Eddy and Maizels, 2008), we hypothesized that strand asymmetries could also be found for the occurrences of G4 around splice sites. In order to study the strand asymmetries patterns of G4 motifs around splice sites across the gene body, for each gene with nine or more exons, we separated the exons of its longest transcript into nine groups: the first four exons, the last four exons and the remaining middle exons. For each of the groups, we calculated G4 motif enrichment at splice sites across both template and non-template strands. We found a pervasive enrichment of G4 motifs across the gene body, however non-template G4 motifs are consistently more highly enriched at 3′ss than template strands (Fig 4.7). In the case of 5'ss, enrichment differences between template and non-template strands are smaller, but it gets bigger towards the gene ends. These findings provide evidence for widespread variation in the topography of G4s in splice junctions; these include the frequency of G4s in the exons and introns flanking the splice site, biases regarding the strand preference, the distance from the splice site and the positioning across the gene body.

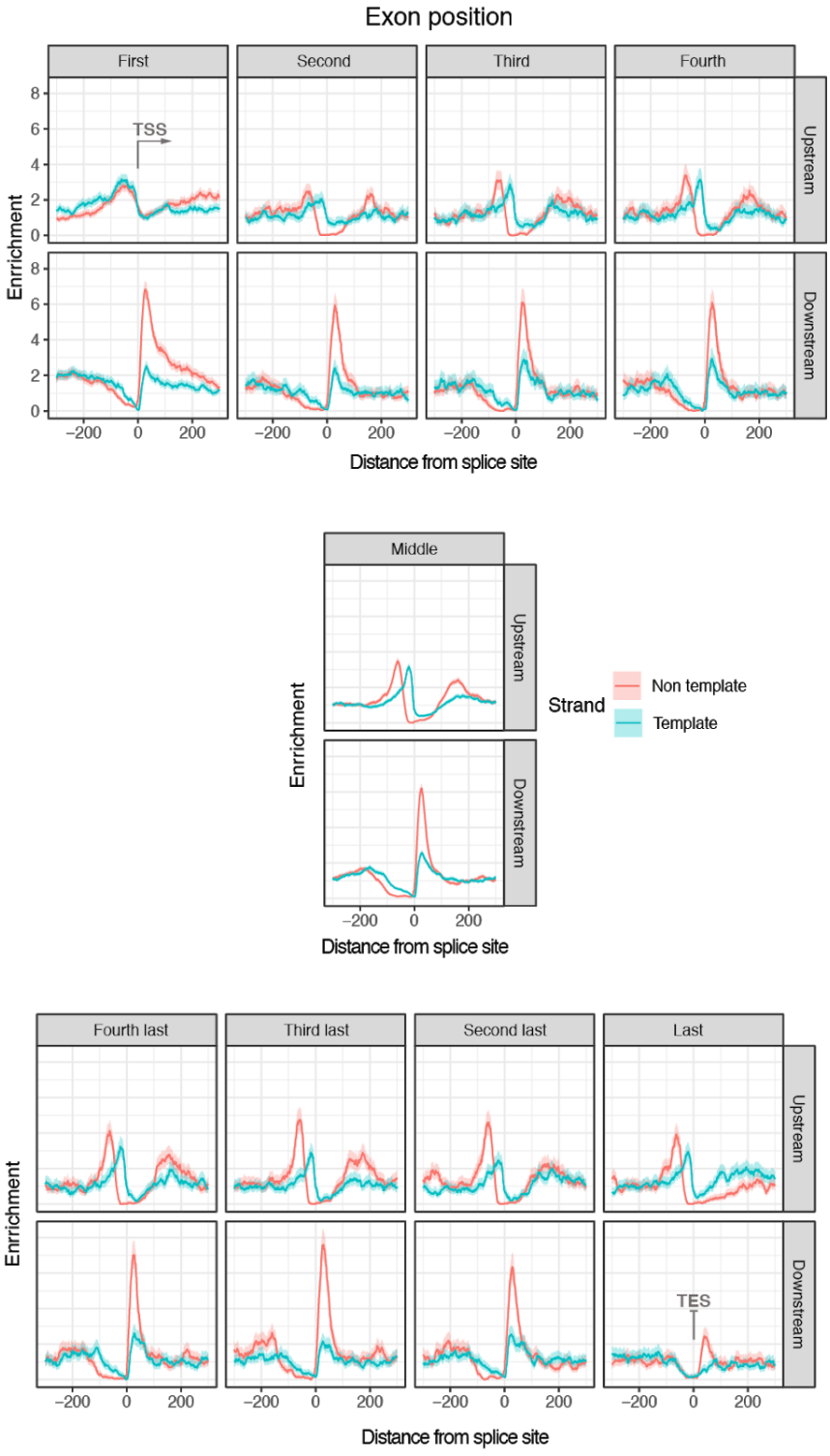

Figure 4.7: Template and non-template splice site G4 enrichment across gene body. G4 motif enrichment relative to splice sites across exons in the gene body for template and non-template strands. Exons were classified according to their gene position. Each panel shows G4 motif enchiment across their upstream or downstream regions of exon across the gene body. The start and end exonic coordinates are centered at 0 x-axis coordinate and they correspond to 3′ss or 5′ss, respectively, except for the first or last exon where transcriptional start site (TSS) and transcriptional end site (TES) are indicated.1.2.2.3 Longer G-runs exhibit higher enrichment around splice sites.

An intramolecular G4 is usually a representation of four or more consecutive G-runs. Yet, fewer consecutive G-runs can also result in G4 formation when they are complemented by additional G-runs from another DNA or RNA molecule (Bhattacharyya et al., 2016; Nasiri et al., 2016). We found minimal to no enrichment for single G-runs at both 5′ss and 3′ss (Fig 4.8). However, for two and three G-runs we observed a 1.39-fold and a 2.10-fold enrichment at the 3′ss and a 1.67-fold and a 2.47-fold enrichment at the 5′ss, which may implicate intermolecular G4s in splice sites. The highest enrichment was observed for four to six G-runs, indicating that intramolecular G4 motifs are more enriched at splice sites than their intermolecular counterparts, in accordance with our earlier findings.

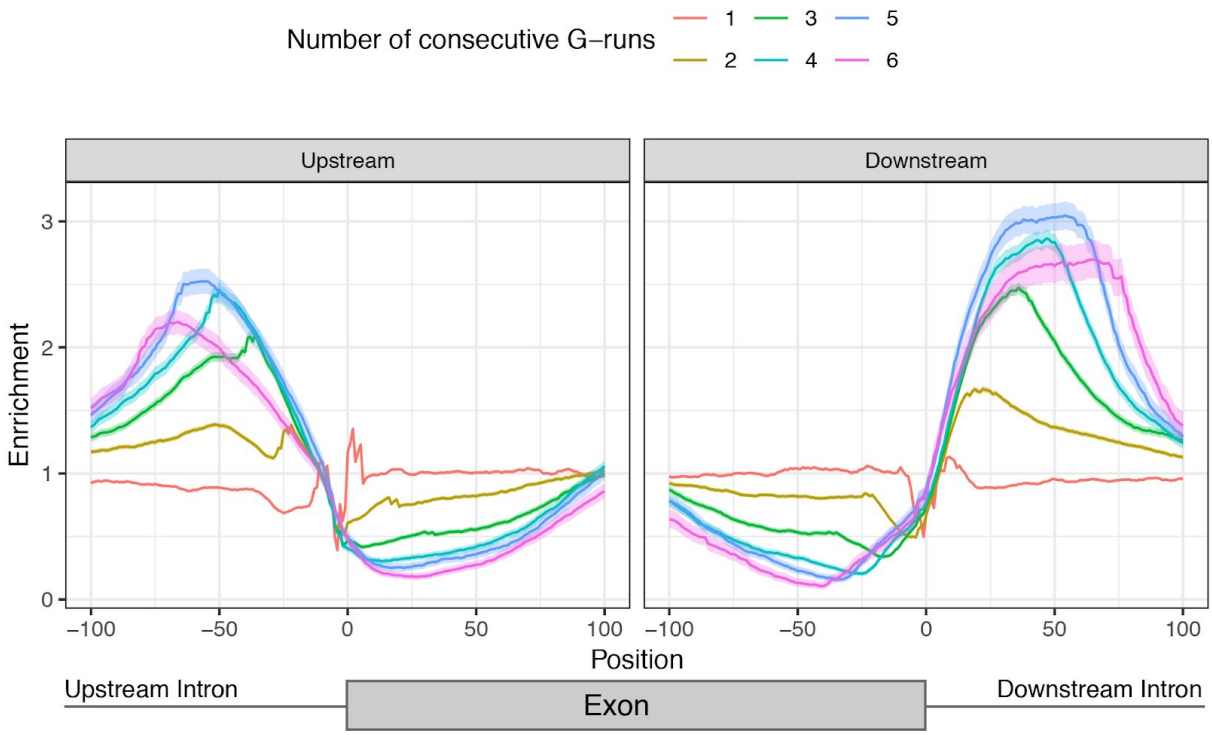

Figure 4.8: Enrichment of G4 of different G-run lengths. Number of consecutive G-runs (G stretches of at least 3 nt separated by 1-7 linker nucleotides) and relative enrichment at the splicing junction. The error bands in A-B represent 0.95 confidence intervals from the binomial error.

4.2.3 Gene architectural features associated with G4-exons

4.2.3.1 G4s are enriched for short introns

The length of introns in metazoans can vary across four orders of magnitude (Sakharkar et al., 2004). We hypothesized that the enrichment patterns of G4s at introns proximal to splicing sites would be associated with intron length. We compared the intron length of splice sites that had a G4 motif within 100 bps in the direction of the intron to the ones that did not have this motif. Consistent with our hypothesis, we found that introns with a G4 at the 3′ss had a median length of 701 nt while introns without a G4 had a median length of 1,618 nt (Figure 4.9a), (Mann Whitney U, p-value<0.001). Similarly, at the 5′ss, introns with a G4 had a median length of 379 nt, whereas introns without a G4 had a median length of 1,629 nt (Mann Whitney U, p-value<0.001). Interestingly, introns in the range of ~45-85 bps were the most enriched for G4s for both the 3′ss and the 5′ss. Moreover, the enrichment of introns in G4s declined rapidly with increased intron length, indicating that they are preferentially found in the subset of short introns (Fig 4.9b-c, Kolmogorov-Smirnov test p-value<0.001). We also investigated the association between splicing strength score and intron length at sites with G4s in the 3′ss and 5′ss and found that the highest enrichment for G4s was in short introns with weak splice site strength (Fig 4.9d-e).

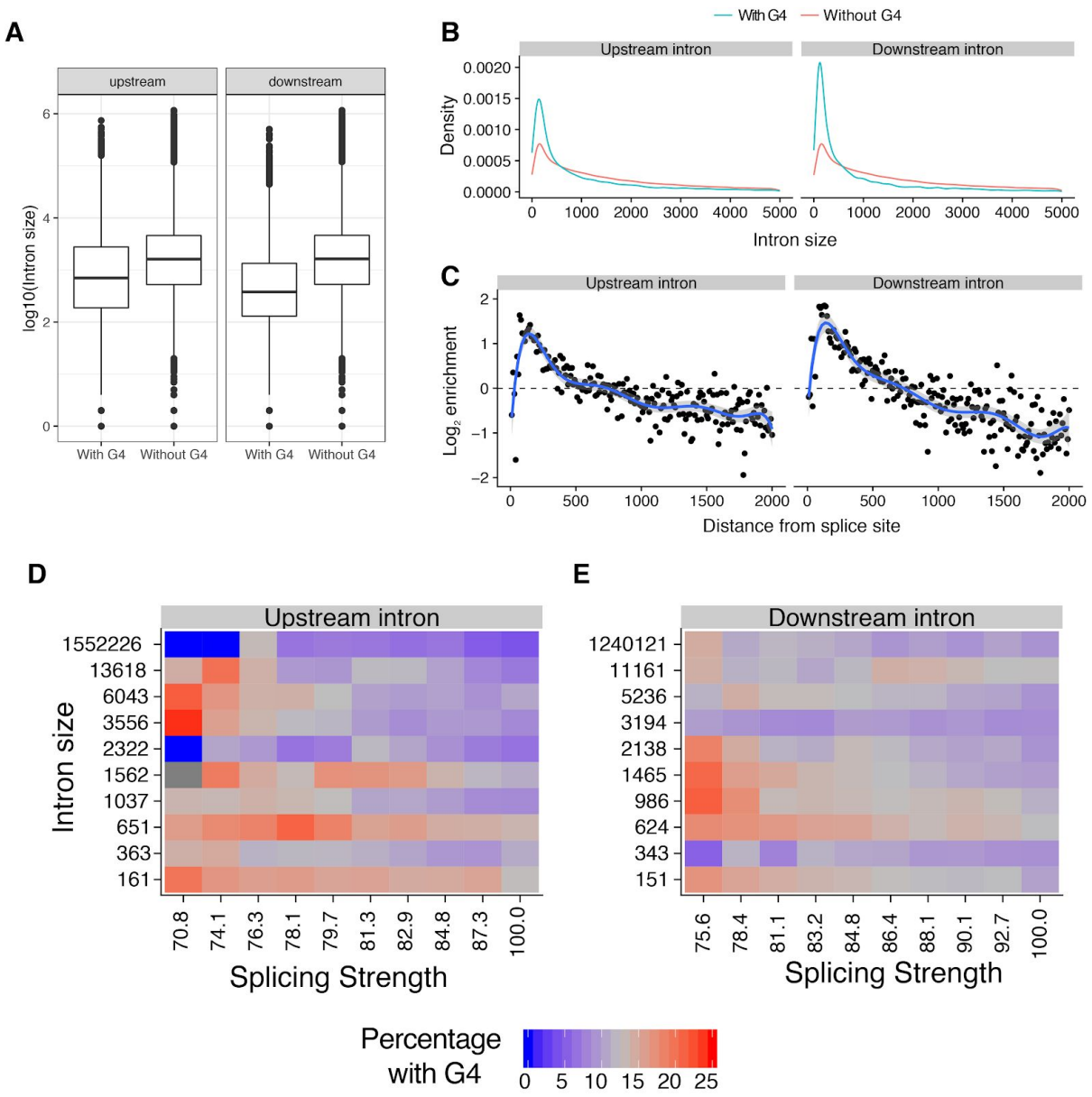

Figure 4.9: G4s are enriched in short introns A. Intron size of the upstream and downstream introns was calculated for groups with or without a G4 within 100 bps of the splice site (Mann-Whitney U p-value <0.001 for both upstream and downstream introns). **B.** Length density distribution of introns upstream and downstream from exons that are flanked by G4s or not (Kolmogorov-Smirnov test p-value<0.001). **C.** Abundance enrichment of intron sizes at upstream and downstream splice sites flanked by G4s. A bin size of 10 bps was used with the blue line representing an eighth degree polynomial model. **D-E.** Heatmap for the relationship between splicing strength score, intron length and G4 presence in a local window of 100 nt within the splice site for the upstream and downstream introns. Red color represents high proportion of splice site regions with G4s, whereas blue color represents depletion of G4s.

4.2.3.1.1 GC-content controlled associations of G4 motifs and intron size

Since short introns are more GC-rich than long introns (Lim and Burge, 2001), we wanted to compare selected groups of introns that have close GC-content distribution (Fig 4.10a). Thus, we divided the intron into two populations, short (<500 nt) and long (>500nt). Then, for different long intron size interval groups, we select groups of short and long introns that have the minimum GC-content difference. By doing this, differences in the fraction of introns containing a G4 in the splice site vicinity (within 100 nt) cannot be attributed to GC-content differences. We found a significantly higher fraction of intron splice sites with a G4 in their vicinity (within 100 nt) for long introns in comparison with short introns, suggesting an inverse enrichment direction when GC-content is controlled (Fig 4.10b).

Moreover, intron size comparisons between template and non-template strands are also not subject to GC-content effect since both strands have the same GC contribution. Thus, to further investigate the relationship regarding the intron length, we separated G4s identified using the consensus motif into non-template and template for both the 5′ss and the 3′ss. At the 3′ss introns showed small but significant differences in length if a G4 was at the non-template or the template strand with medians of 736 nt and 621 nt, respectively (Mann-Whitney U, p-value<1e-21). However, if a G4 was at the non-template strand at the 5′ss the median intron length was 267 bp, whereas if the G4 was at the template strand the median intron length was 539 nt (Mann-Whitney U, p-value<1e-16), displaying more aggravated differences in intron length. Therefore, we conclude that the highest enrichment is found for short introns, on the non-template strand, downstream of the 5′ss.

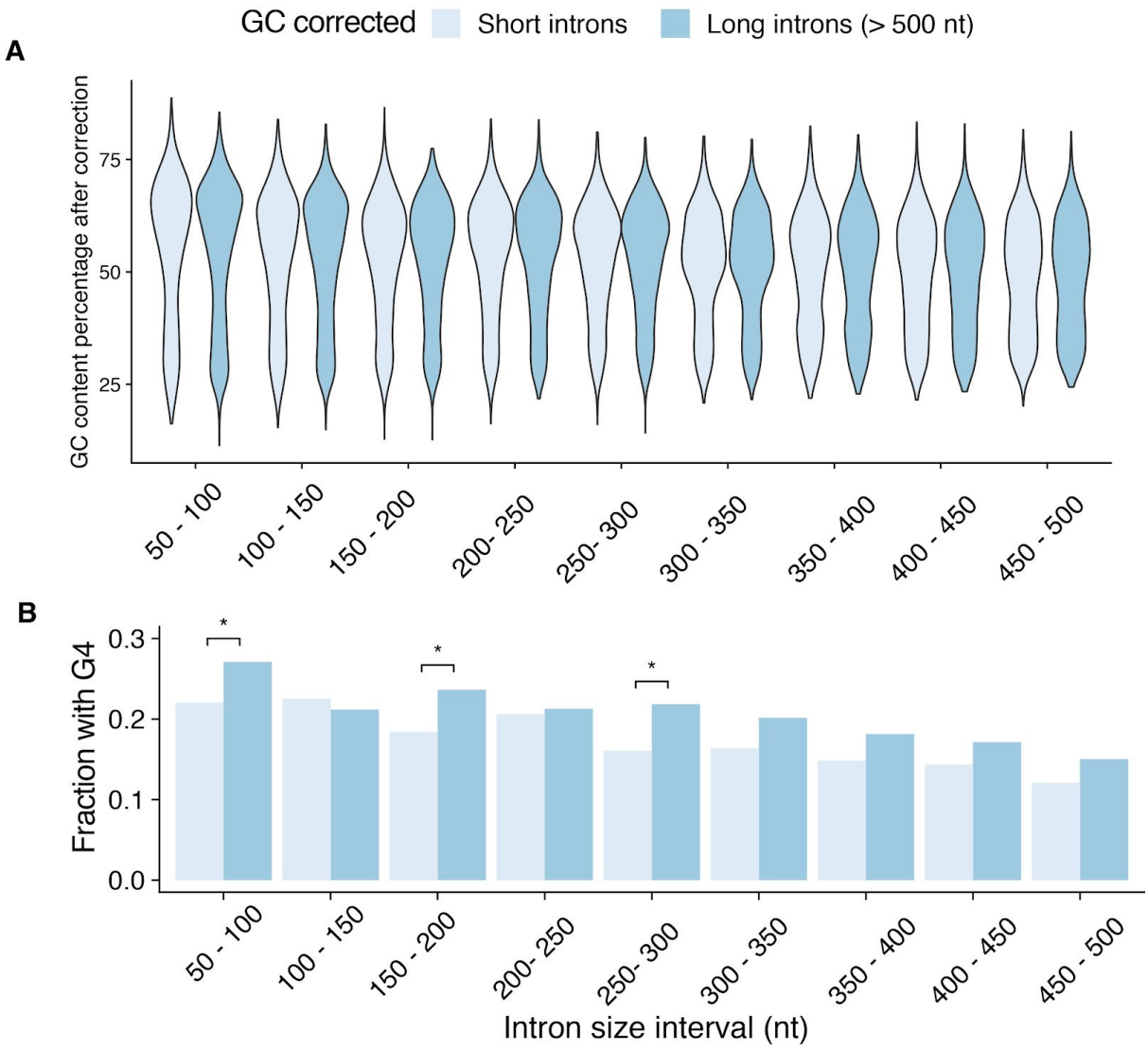

Figure 4.10: G4 enrichment at short intron splice sites is driven by GC-content **A.** GC content distribution across selected groups of short and long introns. Intron size interval refers to the size of small introns. Long introns were defined as introns > 500 bp. **B.** Fraction of splice sites with a G4, controlling for GC content between long and short introns. We use Chi-squared test to evaluate significant differences between short and long introns (* denotes p-values<0.05 after multiple testing corrections).

4.2.3.2 G4 are not enriched in microexons

We also investigated if there is an association between G4s near splice sites and exon length. We do not find a significant association between G4s and exon length at the 3′ss (median exon length without G4s: 124 bp, median exon length with G4s: 123 bp, p-value>0.05, Mann-Whitney U), but we find a significant association for smaller exons near the 5′ss, albeit with a very small magnitude (median exon length without G4s: 127 bp, median exon length with G4: 123 bp, p-value<0.001, Mann-Whitney U). Furthermore, we explored if microexons, defined as exons <30 nt long (Irimia et al., 2014), (Li et al., 2015) had an enrichment for G4s at their splice sites relative to other exons. However, we could not find a higher density of G4s at the introns flanking microexons compared to other exons.

4.2.4 Abundance of G4s at splice sites has emerged during vertebrate evolution

Alternative splicing is a pivotal step of eukaryotic mRNA processing. To understand to what extent splice site regulation by G4s is conserved we considered eleven eukaryotes: *Homo sapiens* (human), *Mus musculus* (mouse), *Sus scrofa* (pig), *Gallus gallus* (chicken), *Danio rerio* (zebrafish), *Caenorhabditis elegans* (nematode) *D. melanogaster* (fruit fly), *Xenopus tropicalis* (frog), *Anolis carolinensis* (lizard), *Saccharomyces cerevisiae* (yeast) and *Arabidopsis thaliana* (flowering plant). *S. cerevisiae* was excluded from further analysis since we could not find any G4s at splice sites and G4s were rare with only 39 occurrences genome-wide. Interestingly, we found that the enrichment pattern of G4 motifs at splice sites was restricted to a subset of vertebrate species, with minimal or no enrichment in fruit fly, *Arabidopsis* and *C. elegans* (Fig 4.11). We observed strong enrichment in chicken, pig, human and mouse, while lizard displayed limited enrichment levels. Surprisingly, *X. tropicalis* and *D. rerio* displayed relative depletion. This suggests that alternative splicing regulation by G4s is found is restricted to mammals and birds, but absent in plants, other tetrapods or fish. However more comprehensive evolutionary analysis are needed to completely discard the presence of G4 enrichment in other organisms.

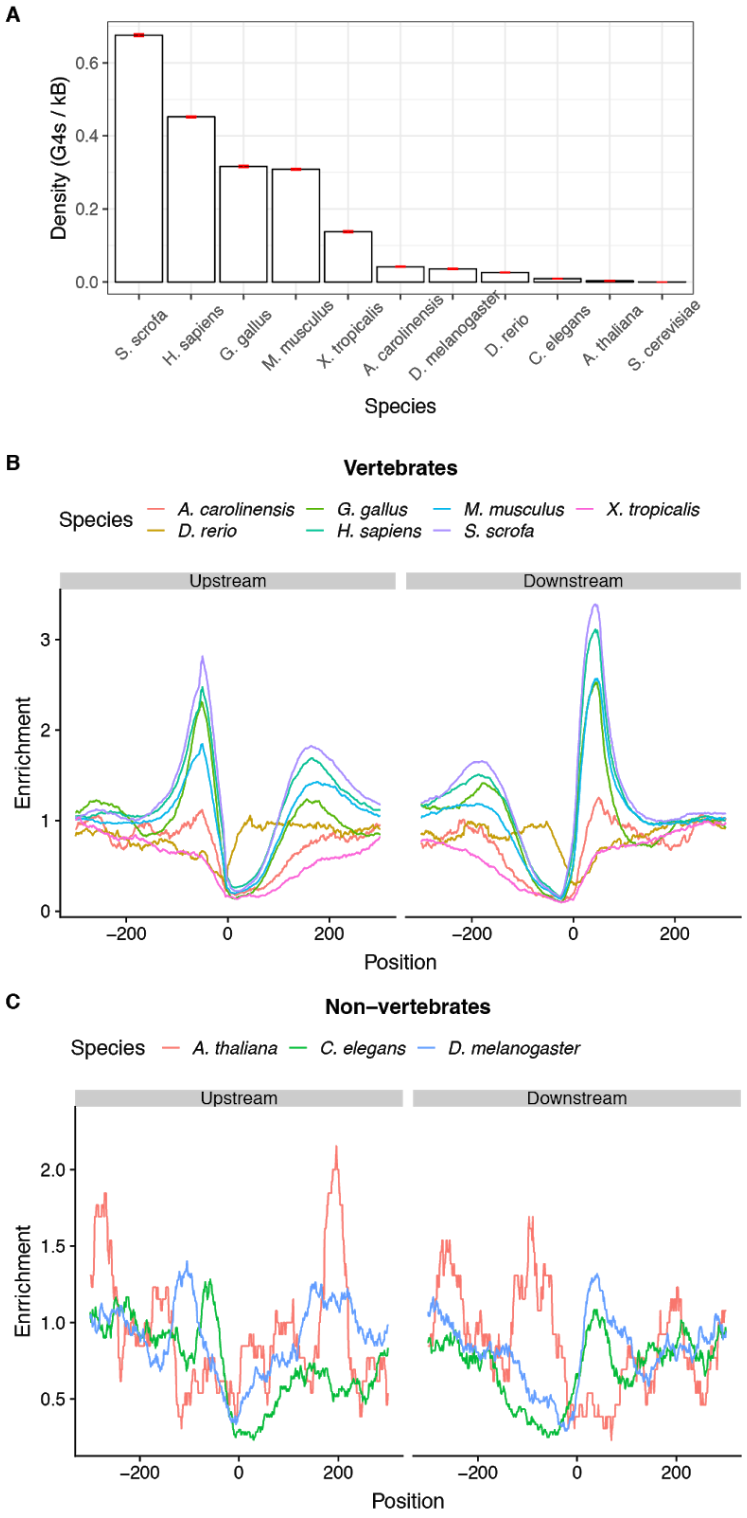

Figure 4.11: G4 motifs are enriched in a subset of vertebrates. A. Density of G4 motifs in a 100 nt window each side across all 5′ / 3′ splice sites of each species. Error bars indicate standard deviation from 1,000-fold bootstrapping with replacement. **B-C.** Enrichment of G4 motifs at splice sites for seven vertebrate (B) and three invertebrate (C) species using the consensus G4 motifs.

Additional support for this conclusion comes from our analysis of G4-seq derived G4 maps generated in PDS and K^+ conditions. These maps are available for multiple model organisms, including three vertebrates (human, mouse and zebrafish) and four non-vertebrate species (nematode, fruit fly, arabidopsis and yeast). Consistent with the analysis based on the primary sequence, we find an acute enrichment of G4s at the 5′ss and 3′ss only in humans and mouse. In particular, we could not find any G4s in the vicinity of splicing junctions for *S. cerevisiae*, there was no enrichment for *D. melanogaster* and *D. rerio*, while we observed a depletion in *A. thaliana* (Fig 4.13).

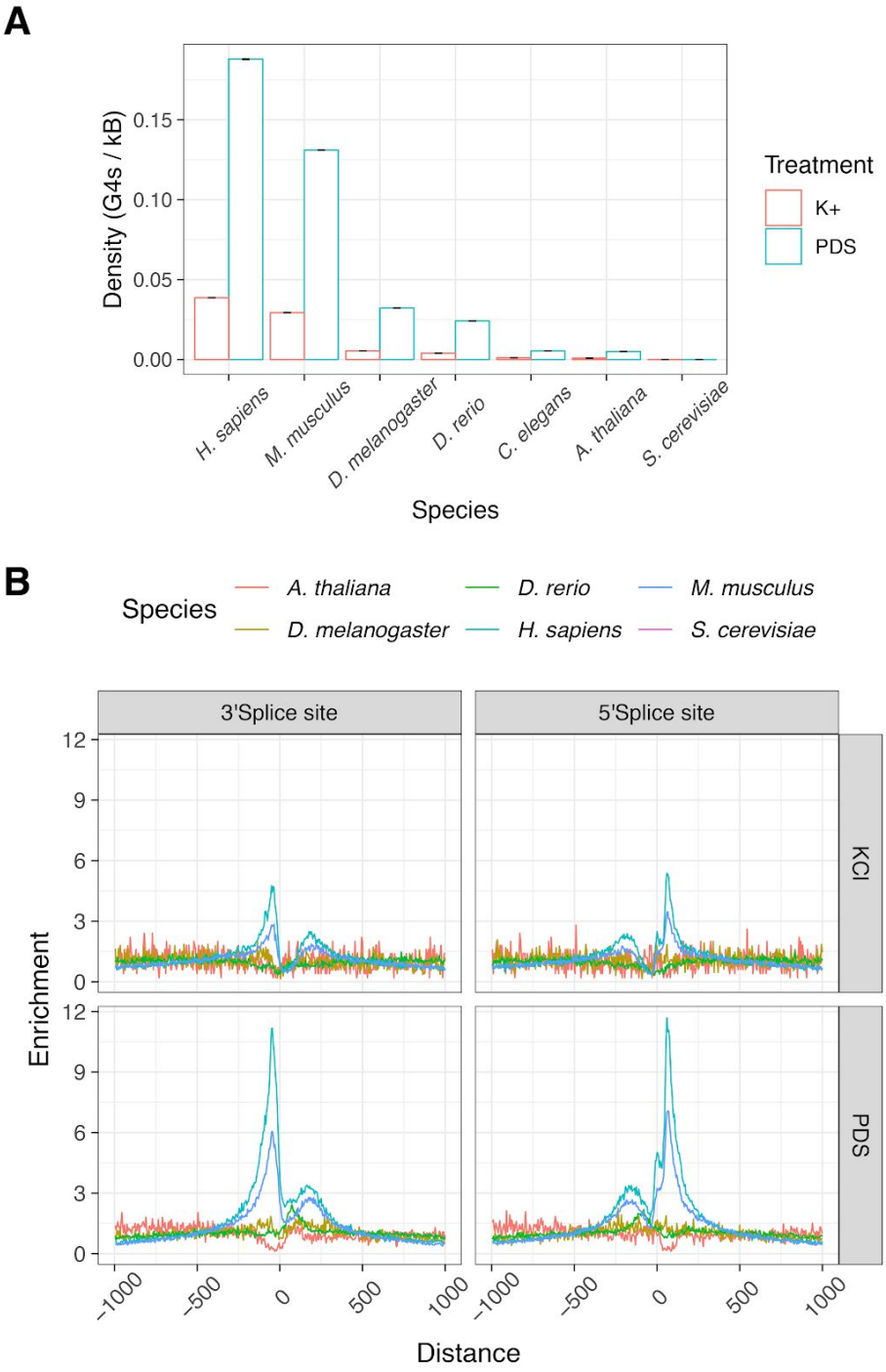

Figure 4.12: Cross specie G4-seq analyses validate findings *in-silico*. A. Enrichment of G4-seq derived G4s at splicing sites at 100 nt splicing site windows in PDS and K^+ treatments. Error bars indicate standard deviation from 1,000-fold bootstrapping with replacement. **B.** Enrichment of G4s at: 5′ / 3′ splice sites across six species for PDS and K^+ treatments.

4.3 Materials and Methods.

4.3.1 Genome and gene annotations processing

We obtained genome assemblies from the UCSC Genome Browser FTP server for eleven organisms: *Homo Sapiens* (hg19), mouse (mm10), *Saccharomyces cerevisiae* (sacCer3), chicken (galGal5), *Drosophila melanogaster* (dm6), zebrafish (danRer11), *Xenopus tropicalis* (xenTro9), *Anolis carolinensis* (anoCar2), *Arabidopsis thaliana* (Tair10) and *Caenorhabditis elegans* (ce10) reference genomes.

We downloaded the Ensembl gene annotation files for the associated genomes from UCSC Table Browser as BED files for each species (Karolchik et al., 2003). Using in-house python scripts we extracted the coordinates of internal exons flanked by canonical splice sites (GT-AG introns) for every species. To calculate the splicing strength scores, we used publicly available positional frequency matrices from the SpliceRack database (Sheth et al., 2006) and previously developed scripts used before for the same purpose (Parada et al., 2014). Splice sites were grouped into quartiles based on their splicing strength score for the downstream analyses to study the distribution of non-B DNA motifs and in particular G4 motifs. The confidence intervals were calculated using “binconf” command from “Hmisc” package in R with default parameters. Mann-Whitney U tests were performed at 100 nt each side in the upstream splice site and at the downstream splice site to compare the splicing strength scores of sites with and without G4s.

4.3.2 Genomic datasets.

4.3.2.1 Non-B DNA motifs.

Identification of each non-B DNA motif was performed using the genome-wide maps in humans and mice provided by (Cer et al., 2013) and processed as described in (Georgakopoulos-Soares et al., 2018). We focused on seven non-B DNA motifs;

inverted repeats, mirror repeats, H-DNA which forms at a subset of mirror repeats with high AG content, G4s, Z-DNA which forms at non-AT alternating purine pyrimidine stretches, short tandem repeats and direct repeats.

Regular expressions were employed to identify genome-wide consecutive G-runs across the human genome, interspersed with loops of up to 7 bps. In total, one to six consecutive G-runs were searched. For each species we generated the genome-wide G4 maps using a regular expression of the consensus G4 motif (G□3N1-7G□3N1-7G□3N1-7G□3). Orientation of G4s and G-runs was performed with respect to template and non-template strands to calculate strand asymmetries at genic regions as previously described for polyN motifs (N being Gs, Cs, Ts and As) in (Georgakopoulos-Soares et al.).

Permuted windows of 100 nt each side of each splice junction were generated using ushuffle (Jiang et al., 2008) correcting for dinucleotide content. The fold enrichment for G4s was calculated as the ratio of the number of motifs found in the real sequences and the median of 1,000 permutations of the set of all real sequences. The corrected enrichment of G4s at 3′ss and 5′ss was calculated as the ratio of the real enrichment of G4s over the background enrichment of G4s at shuffled splice site windows.

To investigate the relationship between non-B DNA motifs or G4-seq peaks and splice sites we generated local windows around the splice sites and measured the distribution of each non-B DNA motif or G4-seq dataset across the window. The enrichment was calculated as the number of occurrences at a position over the median number of occurrences across the window. Regardless of the window size shown in figures, the enrichment was calculated over a window of 1kB. The same approach was used to calculate the enrichment of G4s at splice sites across different species.

The density of G4 consensus motifs or G4-seq derived peaks at local windows was calculated as the number of occurrences of the motif or the peak over the total number of base pairs examined .

4.3.3 G4-seq data

1.4.3.1 G4-seq BedGraph data were obtained from GEO accession code GSE63874 (Chambers et al. 2015) for the human genome and analyzed with bedtools closest command to identify the closest G4 to splice sites and to calculate the distance. The analysis was performed separately for Na+-K+ and Na+-PDS conditions and it was compared to the distribution obtained from the G4 consensus motif. G4-seq BedGraph data for six species, human, mouse, D. melanogaster, C. elegans, A. thaliana and yeast, were obtained from GEO accession code GSE110582 (Marsico et al. 2019) and analyzed using the same genome annotations as those used for the generation of each G4-seq dataset.

Coordinates for internal exons flanked by canonical splice sites (GT-AG introns) were extracted for each species using the Ensembl annotation versions described in (Marsico et al., 2019b) using custom python scripts.

4.3.4 Relationship between G4s and exon / intron length

Introns and exons were grouped based on the presence or absence of G4s within 100 nt each side of the 5′ss and 3′ss and further subdivided into those containing a G4 on the template or on the non-template strand, separately for the 3′ss and the 5′ss. For each of the eight groups we calculated the median length of the intron or exon in a group and performed Mann-Whitney U tests to calculate the significance of the association between length of exons / introns and G4 presence. The R function stat_density was used to plot the length distribution of introns with and without G4s as modelled by a kernel density estimate. Abundance enrichment of intron length in 3′ / 5′ splice sites in relationship with presence of G4s was generated in R using the function geom_smooth in an eighth grade model. Correction of GC content in introns with different length was performed by grouping introns into small introns (<500nt) and large introns (>500nt). Then we calculated the GC content for both groups and for each short intron we selected a long intron with a close GC content value, in such a way that GC distribution across short and long introns groups were nearly identical.

4.3.4.1 G4s and relationship to exon number

For the longest transcript of each gene with nine or more exons we separated exons into 9 groups, the first four exons, the last four exons and the remaining middle exons. To compare the frequency of G4s in splice junctions across the gene body we calculated the distribution of G4s in each exon group relative to the 5′ / 3′ splice sites (S5c). We also calculated the distribution of G4s in each exon group relative to the 5′ / 3′ splice sites separately for the template and non-template strands.

4.3.4.2 Relationship between G4s, splicing strength score and intron length

We calculated the splicing strength score and intron length for the upstream and downstream intron of each exon. We separated introns and splicing strength scores into deciles and calculated the G4 density at each decile, from which we produced two heatmaps displaying the density of G4s as a function of splicing strength score and intron length for the upstream and downstream introns.