rG4-seq reveals widespread formation of G-quadruplex structures in the human transcriptome

Transcriptome-wide profiling of human RNA reveals widespread formation of G-quadruplexes

Chun Kit Kwok[1,2,3], Giovanni Marsico[1,2,3], Aleksandr B. Sahakyan[1,2], Vicki S. Chambers[1,2], Shankar Balasubramanian[1,2]

[1]Department of Chemistry, University of Cambridge, Cambridge, UK.

[2]Cancer Research UK, Cambridge Institute, Cambridge, UK.

[3]These authors contributed equally to this work.

Correspondence should be addressed to S.B. (sb10031@cam.ac.uk).

Keywords: RNA structure, G-quadruplex, transcriptome, gene regulation, rG4-seq, next-generation sequencing

ABSTRACT

We introduce a new transcriptome-wide RNA G-quadruplex (rG4) profiling method called rG4-seq, in which rG4-mediated reverse transcriptase stalling (RTS) is detected by next-generation sequencing. We apply rG4-seq to isolated human HeLa RNA to generate the first global *in vitro* map of rG4 structures, identifying thousands of canonical and non-canonical rG4s. Notably, we show that rG4 formation is related to cytosine (C)-content and alternative RNA structure stability. rG4s are enriched in untranslated regions (UTRs), microRNA target sites, and polyadenylation signals. Structural prediction analysis uncovers rG4-dependent differences in RNA folding. Furthermore, the analysis of rG4s in orthologous genes of eukaryotic species reveals a significant enrichment of rG4-containing transcripts in RNA processing and stability.

RNA folds into diverse structures to control many biological functions and cellular processes[1,2]. Guanine (G)-rich sequences in RNA can assemble into rG4 structures that comprise G-quartets connected by loops (**Supplementary Fig. 1a**). rG4 structures are preferentially stabilised by potassium ions (K^+) but not lithium ions (Li^+)[3]. Recently, rG4s have been demonstrated to exist in human cells[4] and are emerging as important RNA structural motifs involved in many fundamental biological processes including regulation of transcription, RNA processing and translation[5].

Despite rG4's importance in biology, there is so far no experimental method for mapping these structures throughout the transcriptome (**Supplementary Note 1**). Here, we present rG4-seq (RNA G-quadruplex-sequencing), a novel high-throughput approach to profile rG4 structures across the transcriptome at nucleotide resolution (**Fig. 1a**). rG4-seq exploits RTS caused specifically by rG4 formation, either in the presence of the physiological cation K^+ or the rG4-specific ligand, pyrdiostatin (PDS)[6] (**Supplementary Fig. 1b**), to precisely map the location of rG4s.

We first validated our approach using two rG4-containing (positive controls) and two hairpin-containing (negative controls) RNAs (**Supplementary Table 1**) in reverse transcription reactions containing Li^+, K^+, or K^+ plus PDS (K^++PDS) (**Fig. 1a**, see Methods), followed by sequencing. This rG4-seq data showed strong RTS, positioned at the 3'-end of the rG4, only for the positive controls in K^+ and K^++PDS conditions (**Supplementary Fig. 2a**), whereas the negative controls did not exhibit observable RTS (**Supplementary Fig. 2b**). In addition, no or weak RTS was observed in Li^+ condition for the positive controls (**Supplementary Fig. 2a**), confirming rG4-seq is specific for rG4 structures. Moreover, the rG4-seq results were consistent with gel-based analysis of RTS (**Supplementary Fig. 2**), supporting the validity of our approach.

Next, we applied rG4-seq *in vitro* to profile rG4s in purified polyadenylated (polyA) RNA isolated from human HeLa cells. We generated four independent biological replicates with high reproducibility (Pearson Correlation Coefficient (PCC) = 0.978-0.990) and obtained 1.15 billion combined mappable reads covering 17,622 transcripts with Fragments Per Kilobase of transcript per Million of mapped reads (FPKM) $\geq$ 0.5 (**Supplementary Fig. 3**, see Methods). We then compared RTS between K^+ versus Li^+ and K^++PDS versus Li^+ conditions for each biological replicate to generate normalised RTS values for K^+ and K^++PDS conditions (see Methods).

To confidently score an rG4, we required RTS in either K^+ or K^++PDS conditions to have a drop in coverage consistently across all replicates when compared to the control (i.e. Li^+). For this, we used a linear model fitting of the coverage drop values in the four Li^+ replicates versus the four K^+ (or K^++PDS) replicates, followed by ANOVA testing and multiple hypothesis correction (FDR $\leq$ 0.1; see Methods). This approach is similar to RNA-seq differential expression analysis and provides high stringency and confidence in scoring RTS sites (**Supplementary Fig. 4a-c**). Altogether, we scored 3,845 and 13,423 RTS sites for K^+ and K^++PDS conditions, respectively (**Supplementary Tables 2 and 3**).

We found that the majority of our RTS sites occur next to G (or GG, or GGG). Specifically for the K^+ condition, 80.5% and 88.3% of RTS sites have respectively a 'GG' and a 'G' motif within one nucleotide from the stalling site. These values go respectively up to 85.8% and 94.5% when considering three nucleotides from the staling site (**Supplementary Fig. 4d**). Similar values were measured for the K^++PDS condition. We also observed the RTS values are usually higher in the K^++PDS condition versus the K^+ condition (**Supplementary Fig. 4e,f**), supporting that PDS stabilizes rG4s and causes stronger stalling.

These analyses provide a strong indication that the stalling events are specific for rG4s on a transcriptome-wide scale.

We then examined the sequence context of these RTS sites to determine the presence of rG4 structural features using a hierarchical assignment (see Methods). While the consensus sequence for canonical rG4s is $(G_3L_{1-7})_3G_3$, or G_3L_{1-7} for short[7] (**Fig. 1b**), where L denotes loop length in nucleotides, non-canonical rG4s with features such as long-loops, bulges, and 2-quartets (**Fig. 1b**) have also been shown to form under physiological ionic conditions for a few individual candidates[8-10]. In K^+ conditions, 86.4% of the identified RTS sites can be assigned to one of these four structural types (**Fig. 1b,c**) while a small proportion (1.6%) correspond to G-rich sequences (G ≥ 40%) (**Fig. 1c**, see legend). We have validated three of these non-canonical rG4 candidates by *in vitro* selective 2'-hydroxyl acylation[11] experiments, supporting their formation (**Supplementary Fig. 5**). The remaining 12.0% is in close agreement with the FDR threshold of 10%, suggesting that these are likely false positives. Multiple EM for Motif Elicitation **(**MEME) analysis on this 12.0% RTS sites showed that many of the enriched motifs are non-G-rich sequences (**Supplementary Fig. 6**). Thus, 3,383 (88.0%) of the total 3,845 RTS sites (distributed in 2,334 genes) map to rG4 motifs in one of these five classes, indicating the accuracy and robustness of rG4-seq.

In conditions that stabilise rG4 structures (i.e. K^++PDS), we found that the number of RTS sites assigned to any of the five rG4 structural classes increased to 11,367 (84.7% of the total 13,423 RTS sites detected, distributed in 5,817 genes) (**Fig. 1c**). We attribute this mainly to the increased stability caused by PDS on canonical rG4s, and to a larger extent on non-canonical rG4s, which are generally less thermally stable than canonical rG4s[10]. It is possible that PDS might preferentially stabilize certain rG4 structural types, however, strong biases have not been previously observed[12]. MEME analysis of the remaining 15.3% RTS sites showed that many of the enriched motifs are non-G-rich (**Supplementary Fig. 6**). Further, the overlap between K^+ and K^++PDS RTS sites increases from 78.8% to 87.8% when excluding the category "Others" (**Supplementary Fig. 7**). Therefore, to ensure high fidelity in our downstream analyses, we only considered RTS sites that were assigned to one of the five rG4 structural classes, i.e. excluding "Others". Interestingly, we identified scoring rG4s (either in K^+ or K^++PDS) in the 5'UTR of 27 out of the 94 genes (29%) previously reported[13] for having a $(CGG)_4$ motif that showed an effect on translation. Overall, rG4-seq characterises more than ten thousand rG4s in single experiment, which was not previously possible using low-throughput single transcript analysis.

The extensive dataset generated by rG4-seq allows the evaluation of key factors governing rG4 formation in RNA. We assessed how the presence of proximal C bases (90 nt window around the rG4) in RNA affects rG4 formation, since C can base pair with G and compete with rG4 formation. We selected the canonical rG4s (G_3L_{1-7}) detected (**Fig. 1c and Supplementary Fig. 8**) and compared their relative nucleotide content to canonical rG4s without a RTS site (undetected G_3L_{1-7}, i.e. RTS sites with FDR > 0.1; see Methods) (**Supplementary Fig. 8**). We noted that the detected G_3L_{1-7} had a lower C-content than the undetected G_3L_{1-7} (**Supplementary Fig. 9**) for both K^+ and K^++PDS conditions. We also found that C-motifs with 2 or more consecutive Cs yielded a lower ratio (detected G_3L_{1-7} sites

/ undetected G_3L_{1-7} cases) than single C-motifs (**Fig. 1d**). Our analysis indicates that the presence of proximal Cs is disfavoured for rG4 formation, consistent with previous analysis[14] on selected individual transcripts.

We then examined how the presence and stability of alternative RNA structure influences rG4 formation in detected and undetected canonical G_3L_{1-7} cases by using secondary structure prediction program RNA fold[15] (see Methods). We observed that the ΔG (alternative structures) for undetected G_3L_{1-7} cases was significantly lower, i.e. more stable structure, than for the detected G_3L_{1-7} cases in K^+ or K^++PDS ($P = 10^{-84}$ for K^+ and $P = 10^{-56}$ for K^++PDS, Wilcoxon rank-sum test) (**Fig. 1e**), suggesting that stable alternative structures compete with and inhibit rG4 formation.

As an additional validation, we performed *in vitro* selective 2'-hydroxyl acylation experiments to assess RNA structure conformation on a detected and undetected G_3L_{1-7} candidate under Li^+ and K^+ conditions. Our data revealed striking differences in the chemical probing profiles, i.e. change in RNA structure, only for the detected G_3L_{1-7} candidate (**Supplementary Fig. 10**), demonstrating that it forms an rG4 structure under the K^+ condition, and an alternative structure under the Li^+ condition. In contrast, the undetected G_3L_{1-7} candidate formed an alternative structure under both Li^+ and K^+ conditions (**Supplementary Fig. 10**).

We next focused the analysis on the more physiologically relevant condition, i.e. K^+. We assessed the location of detected rG4s in the K^+ condition, and found that rG4s were mostly found in the mRNA (97.7%). Of note, some rG4s detected by rG4-seq were found in transcripts linked to cancers and neurological diseases, such as *PIM1* and *APP* (**Supplementary Fig. 11**), in which the rG4s have previously been shown to negatively regulate translation in cells[16,17]. rG4s have also been predicted in long non-coding RNAs (lncRNAs)[18], and it is notable that we detected rG4s in several lncRNAs, including *MALAT1* and *NEAT1* (**Supplementary Fig. 11**) , demonstrating the first experimental evidence for their formation in this class of RNAs.

We further explored the distribution of rG4s within mRNA transcripts and discovered that of the 3,383 total rG4s observed in K^+ conditions, 3′-UTRs contained the majority (2,086, 61.7%) compared to 5′-UTRs (540, 16.0%) or coding sequence (CDS) (697, 20.6%) (**Fig. 2a**). Some rG4s overlap multiple transcript regions and have been assigned to both categories (17 rG4s at 3′-UTR and CDS; 27 rG4s at 5′-UTR and CDS). The remaining 104 scoring rG4s were located in 57 non-coding transcripts, such as lncRNAs and pseudogenes (**Supplementary Table 2**). To take into account length differences in each region, we calculated rG4 density after normalising by the total length of each region, and found rG4s to be 4-5 times more enriched in UTRs than CDS (**Fig. 2b**). When the rG4 distribution was analysed by partitioning each region into bins (region length normalisation), we observed no apparent positional bias in UTRs (**Fig. 2c**, see Methods).

As UTRs are known to contain RNA cis-regulatory elements controlling diverse biological processes, our findings suggest rG4s may have regulatory roles for many

transcripts. To further explore this hypothesis, we first investigated the relationship between rG4 sites and microRNA (miRNA) target sites. We measured the distance from each rG4 to the closest miRNA target sites and found that rG4s were significantly enriched near miRNA target sites compared to a control analysis, in which the rG4s were randomly shuffled (**Fig. 2d**; see Methods). For example, 36.7% of rG4s exhibited a miRNA target site within 100 nt, in contrast to 20.9% by chance ($P = 10^{-37}$, Chi-squared test for proportions) (**Fig. 2d**). Significant enrichments were also observed when we considered separately the rG4s downstream ($P = 10^{-45}$) or upstream ($P = 10^{-13}$) of the miRNA target sites (**Fig. 2d**), showing that rG4s are enriched on both sides. Measuring this at a shorter distance (e.g. 50 nt) also revealed significant association (25.2% versus 15.0% random, $P = 10^{-20}$). Our findings suggest a regulatory role of rG4s in the miRNA pathway, in line with recent studies showing that rG4s near a miRNA target site can regulate miRNA binding accessibility[19], or bind to proteins that regulate AGO2 association[20].

We next explored the association between rG4 sites and polyadenylation signals (PASs) (see Methods). We found that rG4s were significantly enriched near PASs (within 100 nt) when compared to random reshuffling ($P = 10^{-3}$) (**Fig. 2e**). Similarly to miRNA target sites, we also observed significant enrichment when we separately assessed rG4s downstream ($P = 10^{-6}$) or upstream ($P = 10^{-5}$) of the PASs (**Fig. 2e**). rG4s near PASs have been shown to regulate alternative polyadenylation in *LRP5* and *FXR1*[21], and together with our transcriptome-wide data, these suggest a general regulatory role of rG4s in the polyadenylation pathway.

Little is known about the impact of rG4 formation on RNA folding, therefore we investigated the extent to which the presence of the rG4s detected by rG4-seq would affect RNA structure, using RNAfold[15] on 3,319 unique sequences of 250 nt around rG4s (see Methods). To this end, we constrained nucleotides within each identified rG4s to prevent Watson-Crick base-pairing (rG4-constrained prediction) and compared the resultant structures to the unconstrained cases by using positive predictive values (PPV) analysis[22,23] (see Methods). If the presence of rG4 would affect only the local RNA secondary structure, high structural similarity (i.e. PPV values) would be expected since the rG4-constrained region is approximately 10% of the total sequence length. Notably, we found that the majority of secondary structures differed extensively in presence of the rG4-constraint (median PPV = 46.75%) (**Fig. 3a**), suggesting that factoring in the rG4 motif leads to different conformations of RNA and affects the longer-range folding of distal domains of the RNA structure (**Fig. 3b and Supplementary Fig. 12**). In some cases it may be possible that two (or more) RNA conformations coexist, as shown by studies on hairpin-rG4 systems in selected transcripts[11,24], whereby perturbing the conformational equilibrium regulates biological functions[24]. We speculate that hairpin-rG4 structure transitions maybe prevalent in human mRNAs and contribute to gene regulation.

Gene ontology (GO)[25] analysis for all rG4-containing transcripts reveals associations with the regulation of biosynthetic processes, transcription, and chromatin organisation. To provide further insight, we examined the occurrence of rG4s in the complete set of 68 eukaryotic species with available reference sequences in Ensembl (**Supplementary Fig. 13**).

While specifically considering the matching loci of orthologous genes (see Methods), we noted the presence of a group of rG4s with a strong cross-species occurrence (CSO) spanning a wide range of species (**Fig. 3c**, compare strong and average CSO groups). GO analysis of the unique genes in either group (**Fig. 3d, Supplementary Table 4**) revealed 54 significantly enriched terms exclusive to the rG4s from the strong CSO group (**Fig. 3e, Supplementary Fig. 14**). Notably, among those were terms related to RNA processing, regulation of transcription, and RNA stability (**Fig. 3e**). In a broad sense, we suggest that rG4-seq can be applied both to generate and test hypotheses of rG4 candidates that may have importance in biology.

In this work, we develop the first high-throughput experimental approach for mapping rG4s transcriptome-wide at nucleotide resolution. Application of rG4-seq *in vitro* to the human HeLa RNA reveals that the rG4 is a pervasive RNA secondary structure in the human transcriptome and might be a regulatory element for processes that include miRNA-mediated gene regulation and alternative polyadenylation. The significant influence of rG4s on predicted RNA secondary structures in transcripts suggests rG4s should be considered in any RNA structure map (RNA structurome[2]) and in the dynamic interconversion of RNA structures. The repertoire of rG4s identified in this study provides a valuable resource for those interested in RNA and gene regulation. Given the growing interest in rG4s and their relevance to gene regulation and diseases, we anticipate that rG4-seq and the findings we report herein will help stimulate development of *in vivo* rG4 transcriptome-wide profiling methods (**Supplementary Note 2**) and elucidate the functions of this important RNA secondary structure in biology.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. NCBI's GEO repository under the accession number GSE77282 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE77282)

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGEMENTS

This study is supported by an European Research Council Advanced Grant No. 339778 (S.B.), a CASE studentship from Biotechnology and Biological Sciences Research Council (BBSRC) and Illumina® BB/I015477/1 (V.S.C), a Herchel Smith Fellowship (A.B.S.), and

some support from Croucher Foundation (C.K.K). S.B. is a Senior Investigator of the Wellcome Trust. We thank members of the Balasubramanian laboratory for comments.

AUTHOR CONTRIBUTIONS

C.K.K., G.M., A.B.S., V.S.C., S.B. designed the experiments, C.K.K., G.M., and A.B.S. performed the experiments and data analysis. C.K.K., G.M., A.B.S., V.S.C., and S.B. interpreted the results and co-wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1 Wan, Y., Kertesz, M., Spitale, R. C., Segal, E. & Chang, H. Y. Understanding the transcriptome through RNA structure. *Nature Reviews Genetics* **12**, 641-655 (2011).
2 Kwok, C. K., Tang, Y., Assmann, S. M. & Bevilacqua, P. C. The RNA structurome: transcriptome-wide structure probing with next-generation sequencing. *Trends Biochem. Sci.* **40**, 221-232 (2015).
3 Neidle, S. & Balasubramanian, S. *Quadruplex nucleic acids*. Vol. 7 (Royal Society of Chemistry. Cambridge, UK., 2006).
4 Biffi, G., Di Antonio, M., Tannahill, D. & Balasubramanian, S. Visualization and selective chemical targeting of RNA G-quadruplex structures in the cytoplasm of human cells. *Nat. Chem.* **6**, 75-80 (2014).
5 Millevoi, S., Moine, H. & Vagner, S. G-quadruplexes in RNA biology. *WIREs RNA* **3**, 495-507 (2012).
6 Rodriguez, R. *et al.* A novel small molecule that alters shelterin integrity and triggers a DNA-damage response at telomeres. *J. Am. Chem. Soc.* **130**, 15758-15759 (2008).
7 Huppert, J. L., Bugaut, A., Kumari, S. & Balasubramanian, S. G-quadruplexes: the beginning and end of UTRs. *Nucleic Acids Res.* **36**, 6260-6268 (2008).
8 Jodoin, R. *et al.* The folding of 5'-UTR human G-quadruplexes possessing a long central loop. *RNA* **20**, 1129-1141 (2014).
9 Martadinata, H. & Phan, A. T. Formation of a stacked dimeric G-quadruplex containing bulges by the 5'-terminal region of human telomerase RNA (hTERC). *Biochemistry* **53**, 1595-1600 (2014).
10 Pandey, S., Agarwala, P. & Maiti, S. Effect of loops and G-quartets on the stability of RNA G-quadruplexes. *J. Phys. Chem. B* **117**, 6896-6905 (2013).
11 Kwok, C. K., Sahakyan, A. B. & Balasubramanian, S. Structural analysis using SHALiPE to reveal RNA G-quadruplex formation in human precursor microRNA. *Angew. Chem. Int. Ed.* **DOI: 10.1002/anie.201603562** (2016).
12 Chambers, V. S. *et al.* High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nat. Biotechnol.* **33**, 877-881 (2015).

13 Wolfe, A. L. *et al.* RNA G-quadruplexes cause eIF4A-dependent oncogene translation in cancer. *Nature* **513**, 65-70 (2014).
14 Beaudoin, J. D., Jodoin, R. & Perreault, J. P. New scoring system to identify RNA G-quadruplex folding. *Nucleic Acids Res.* **42**, 1209-1223 (2014).
15 Lorenz, R. *et al.* ViennaRNA package 2.0. *Algorithms Mol Biol* **6**, 26 (2011).
16 Arora, A. & Suess, B. An RNA G-quadruplex in the 3' UTR of the proto-oncogene PIM1 represses translation. *RNA Biol.* **8**, 802-805 (2011).
17 Crenshaw, E. *et al.* Amyloid precursor protein translation is regulated by a 3'UTR guanine quadruplex. *PLoS One* **10**, e0143160 (2015).
18 Jayaraj, G. G., Pandey, S., Scaria, V. & Maiti, S. Potential G-quadruplexes in the human long non-coding transcriptome. *RNA Biol.* **9**, 81-86 (2012).
19 Stefanovic, S., Bassell, G. J. & Mihailescu, M. R. G quadruplex RNA structures in PSD-95 mRNA: potential regulators of miR-125a seed binding site accessibility. *RNA* **21**, 48-60 (2015).
20 Kenny, P. J. *et al.* MOV10 and FMRP regulate AGO2 association with microRNA recognition elements. *Cell Rep* **9**, 1729-1741 (2014).
21 Beaudoin, J. D. & Perreault, J. P. Exploring mRNA 3'-UTR G-quadruplexes: evidence of roles in both alternative polyadenylation and mRNA shortening. *Nucleic Acids Res.* **41**, 5898-5911 (2013).
22 Reuter, J. S. & Mathews, D. H. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* **11**, 129 (2010).
23 Ding, Y. *et al. In vivo* genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature* **505**, 696-700 (2014).
24 Pandey, S., Agarwala, P., Jayaraj, G. G., Gargallo, R. & Maiti, S. The RNA stem-loop to G-quadruplex equilibrium controls mature microRNA production inside the cell. *Biochemistry* **54**, 7067-7078 (2015).
25 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29 (2000).

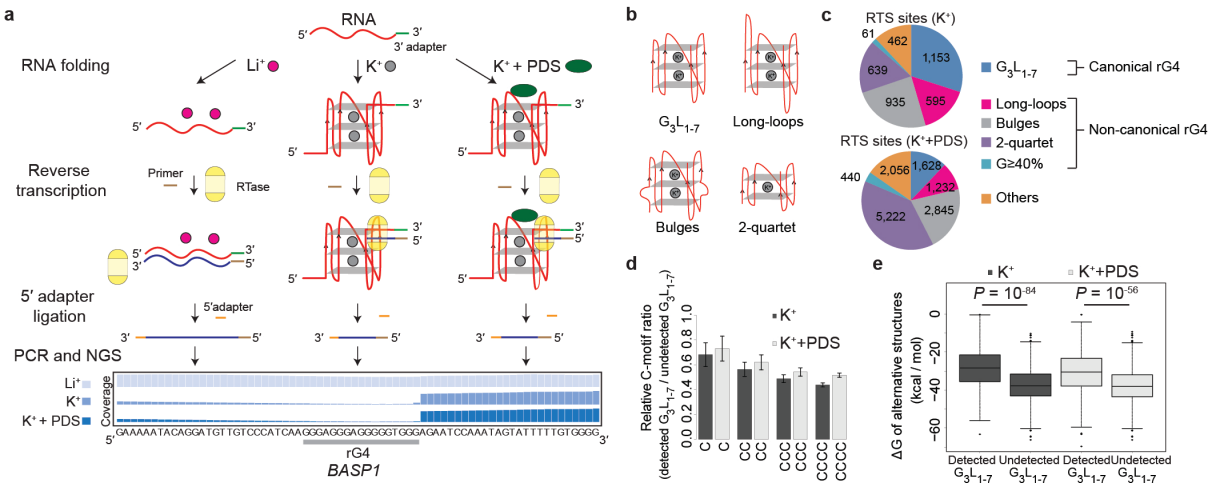


Figure 1 | rG4-seq is a novel transcriptome-wide method to profile rG4 structure and reveals unique rG4 structural features. (**a**) Overview of rG4-seq. RNA (red) is ligated to 3' adapter (green), followed by RNA folding under different conditions (Li^+, K^+, or K^++PDS). rG4 induces reverse transcriptase (RTase) stalling during reverse transcription, leading to cDNA fragments of different lengths (purple). The cDNAs are ligated to 5' adapter (orange), followed by PCR and next generation sequencing (NGS). *BASP1* (chr5:17,276,185-17,276,254) is shown as an example. For K^+ and K^++PDS conditions, the drop in coverage (from 3' to 5' direction) is caused by rG4 (grey box) formation, whereas in Li^+ condition, the coverage is generally uniform since no rG4 is formed. (**b**) Illustration of rG4 structural subclasses. $G_3L_{1\text{-}7}$, canonical rG4s with loop length 1-7 nt; long-loops, rG4s with loop length >7 nt; bulges, rG4s with bulges within the G-tract; 2-quartet, rG4s with 4-tracts of two consecutive Gs. (**c**) RTS sites in different rG4 categories for K^+ and K^++PDS (Methods). G≥40%: sequences with at least 40% G-content not falling into the categories in (**b**). (**d**) Relative C-motifs ratio on detected $G_3L_{1\text{-}7}$ versus undetected $G_3L_{1\text{-}7}$ for K^+ and K^++PDS. Errors bars: standard deviations. (**e**) ΔG of alternative structures on detected $G_3L_{1\text{-}7}$ and undetected $G_3L_{1\text{-}7}$ for K^+ and K^++PDS. Boxes: 25^{th}-75^{th} percentile with median marked; whiskers: 5^{th}-95^{th} percentile.

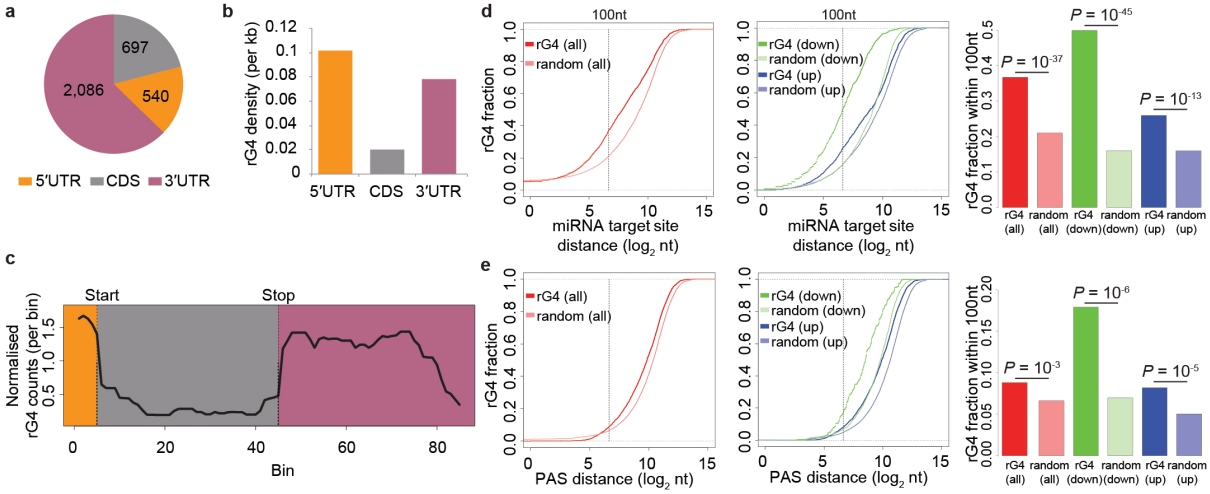


Figure 2 | rG4-seq uncovers rG4 enrichment in UTRs of mRNA and an association with miRNA target sites and PASs. (**a**) The number of rG4s in different regions of mRNA. (**b**) The rG4 density per kilobase (kb) in different regions of mRNA. (**c**) The average normalised rG4 counts along a transcript standardised to the same length and equally binned. (**d**) Association of rG4s with miRNA targeting sites. rG4s are significantly enriched in microRNA target sites, more than expected by random chance alone. Colour legend for curves and bars: red = all rG4s; green = rG4s downstream of sites; blue = rG4s upstream of sites. (**e**) Association of rG4s with PAS, colours as in (**e**).

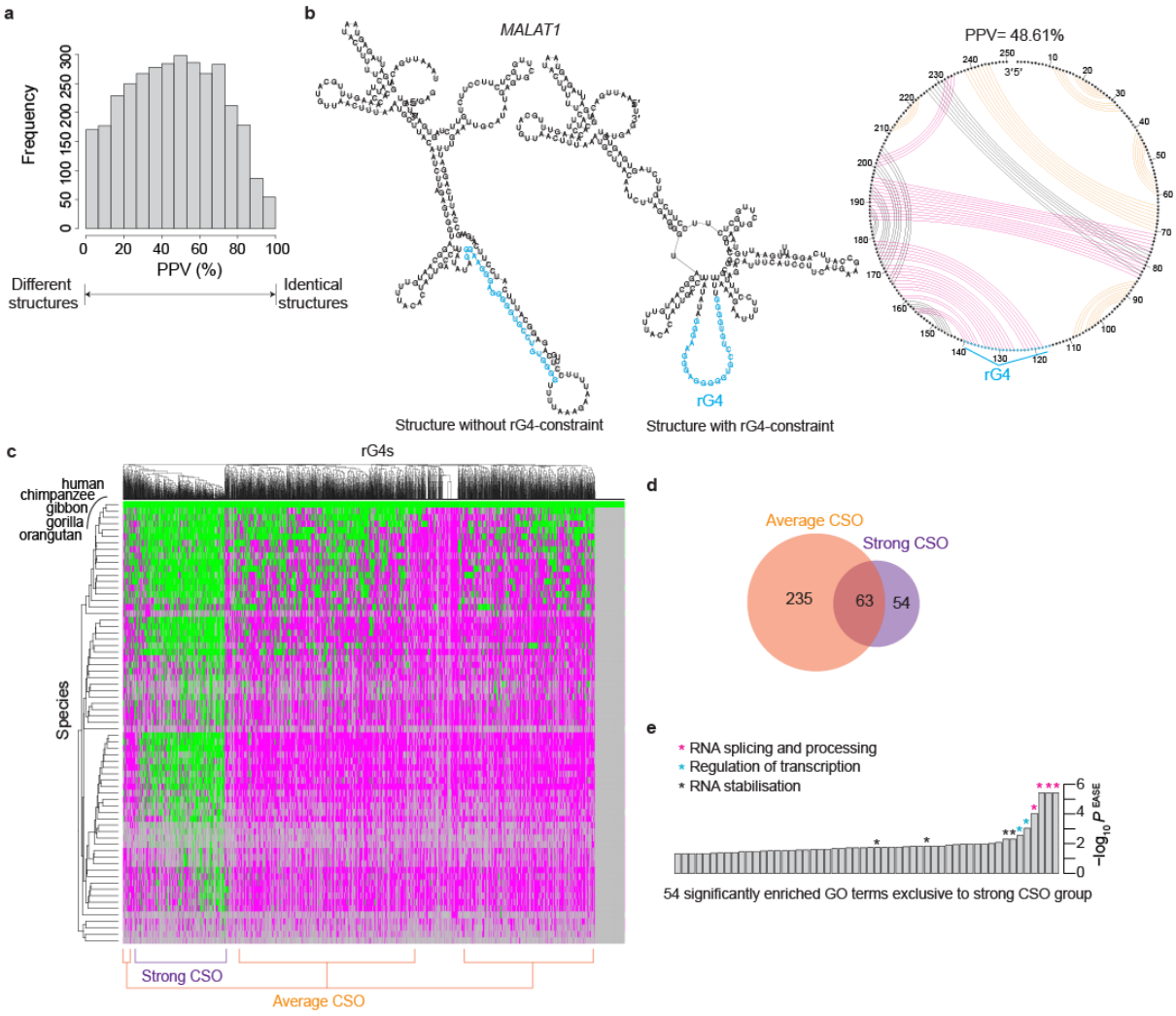


Figure 3 | rG4-seq identifies rG4-dependent differences in RNA conformations and reveals functional associations for the rG4-containing transcripts. (**a**) Distribution of PPV for 3,319 RNA structures with and without rG4 constraints. Lower PPV indicates more structural differences. Median PPV = 46.75% (**b**) Representative examples of RNA structure with and without rG4-constraint (*MALAT1*, chr11:65,271,580-65,271,829; PPV = 48.61%). rG4 is highlighted in blue. In the circle plot, orange is base pairs present in both structures. Black and magenta, base pairs present only in structure with or without rG4-constraint, respectively. (**c**) Heatmap showing the occurrence of human rG4 analogues (columns) in 68 eukaryotic species (rows). First row: data for human (all green). Other rows: species where a specific rG4 is present (green) or absent (magenta) in the matching aligned locus of an ortholog, or the ortholog is not found (grey). Groups of rG4 cross-species occurrence (CSO): strong CSO - rG4s present across the orthologs of most species; average CSO - rG4s present in orthologs of only some species (**Supplementary Fig. 13**). (**d**) Venn diagram showing the number of GO terms significantly enriched for both average and strong CSO groups. (**e**) The 54 GO terms exclusive to the strong CSO group of rG4s, shown against their significance level ($-log_{10}P^{EASE}$). Three frequent categories of related GO terms are outlined.

ONLINE METHODS

RNA preparation. RNAs for gel-based RTS assay and SHAPE were transcribed by T7 *in vitro* transcription using New England Biolabs (NEB) HiScribe T7 High Yield RNA Synthesis Kit. The RNAs were purified by 15% denaturing polyacrylamide gel, crushed and soaked with 1× $TELi_{800}$ overnight according to previous literature[26], and subsequently cleaned and concentrated with RNA clean and concentrator (Zymo research).

Gel-based RTS assay. Three pmol of RNA in 4.5 µl nuclease-free water was mixed with 1 µl of 5 µM Cy5-labeled transcript-specific reverse primer and 3 µl of 5× reverse transcription reaction buffer (final conc. 20 mM Tris, pH 7.5, 4 mM $MgCl_2$, 1 mM DTT, 0.5 mM dNTPs, 150 mM LiCl or 150 mM KCl). The mixture was heated at 95°C for 1.5 min and cooled at 4°C for 1.5 min, followed by 37°C for 15 min for system equilibration. Note that the RTS assay and rG4-seq (below) were performed under physiological relevant [K^+], [$Mg2^+$], pH, and temperature, which are close to those found in cells. At the beginning of the 37°C incubation, 1 µl of nuclease-free water or 50 µM of PDS was added to the reaction and mix thoroughly. For dideoxy sequencing, 1 µl of 10 mM corresponding dideoxy nucleotide was added instead. After the 15 min incubation, 0.5 µl of Superscript III (200U/µL) was added and the reverse transcription was carried out at 37°C for 50 min, followed by treatment of NaOH at 95°C for 10 min. Next, 10 µl of 2× stopping solution which contains 20 mM Tris, pH 7.5, 20 mM EDTA, 94% deionized formamide was added to the reaction mixture. The cDNAs were fractionated by 8% denaturing polyacrylamide gel.

rG4-seq library preparation. Authenticated human *HeLa* cells with no mycoplasma contamination were cultured in DMEM media supplemented with 10% fetal bovine serum (Sigma) as previously described[27]. The cells were pelleted and total RNA was extracted using Qiagen RNeasy Plus Mini Kit following manufacturer's protocol. Genomic DNA was removed during the RNA extraction process by gDNA eliminator columns provided in Qiagen RNeasy Plus Mini Kit.

To enrich for polyA RNAs, 300 µg of total RNA was used per polyA purist kit (Ambion) and yield ~1.5 µg after two rounds of polyA selection. RNA random fragmentation was performed in 40 mM Tris-HCl pH 8.2, 100 mM LiCl, 30 mM $MgCl_2$ at 95°C for 90 s to yield average fragment size of ~250 nucleotides, followed by RNA clean and concentrator (Zymo research). 3' dephosphorylation was performed using 8 µl sample, 1 µl 10× T4 PNK buffer, 1 µl T4 PNK enzyme (NEB) with no ATP added at 37°C for 30 min. Next, 3'adapter ligation was conducted by adding 10 µl sample from above, 1 µl of 50 µM 3'rApp adapter (5'-/5rApp/AGATCGGAAGAGCACACGTCTG/3SpC3/-3'), 1 µl 10× T4 RNA ligase buffer, 6 µl PEG8000, and 2 µl T4 RNA ligase 2 K227Q (NEB) at 25°C for an hour, followed by RNA clean and concentrator. The sample was then divided into three parts for 150 mM Li^+, 150 mM K^+, and 150 mM K^++ 5 µM PDS for reverse transcription with 1 µl of 10 µM unlabelled reverse primer (5'-CAGACGTGTGCTCTTCCGATCT-3') and followed as described above (see gel-based RTS assay), except with in 20 µl. After reverse transcription, the cDNAs were purified by 10% denaturing TBE gel and the size ~35-500nt was sliced. The gel was crushed and soaked in 1×TEN_{250} and incubated at 80°C for 30 min, followed by RNA clean and

concentrator. To the purified cDNAs (8 µl) was added with 1 µl 50 µM 5'adapter (5'/5Phos/AGATCGGAAGAGCGTCGTGTAGCTCTTCCGATCTNNNNNN/3SpC3/3'). The sample was heated at 95°C for 3min, cooled to room temperature, and 10 µl of 2× Quick T4 ligase buffer and 1 µl Quick T4 DNA ligase (NEB) was added and reacted at 20°C overnight. The ligated cDNAs were purified by 10% denaturing TBE gel and the size 75-500nt was sliced, followed by gel extraction step as described above. Next, PCR reaction (25 µl) was performed using 95°C: 3 min, 9 cycles:(98°C: 20 s, 65°C: 15 s, 72°C: 40 s), 72°C: 1 min, 1 µl 10 µM forward primer (5' AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGA TCT 3') and 1 µl 10 µM reverse primer (e.g. index 2) (5' CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTCAGACGTGTGCT CTTCCGATCT 3'), 10.5 µl template and 12.5 µl 2× KAPA HiFi readymix. The amplified libraries were purified with 1.8% agarose gel, and size 150-500 bp was sliced and extracted with Genejet gel extraction kit (Thermo Scientific). The purified libraries were qPCR with KAPA Universal Quant Kit, and subjected for next generation sequencing on NextSeq500 machine for 1 × 150 bp cycle run. Libraries for control RNAs were prepared separately using the same protocol described above, beginning from the step of reverse transcription.

Sequencing data mapping and alignment. Sequencing data were pre-processed by the trim galore software (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) for removal of Illumina sequencing adapters and low quality read tails. Trimmed data were then aligned to the human reference genome version *hg19* by using the *tophat2* software[28] (https://ccb.jhu.edu/software/tophat/index.shtml). The human genome sequence and gene annotation files for the alignment have been downloaded from the Illumina iGenomes support website (https://support.illumina.com/sequencing/sequencing_software/igenome.html). Aligned read with mapping quality below 30 were removed using the samtools[29] (http://samtools.sourceforge.net). The coverage bedGraph files were calculated using the bedtools (http://bedtools.readthedocs.org/).

Data normalization and scoring pipeline for rG4-seq. To identify which genomic regions were affected by the formation of G-quadruplex, i.e., which regions show a drop in coverage indicating a stalling during the reverse transcription reaction, the following analysis was performed.

For each biological replicate separately:

1) exons of the same gene were merged (*bedtools merge*) and the single base coverage was calculated (*bedtools genomecov*), yielding a 1-dimensional coverage signal;
2) the coverage signal was convolved with a step-like filter of order 10 to highlight drops in signal at each base (R^{31} function *filter*, with the *convolution* option, coefficients [1 1 1 1 1 1 1 1 1 1 0 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1]);
3) the coverage signal was convolved with a different step-like filter of order 10 to normalize for the total coverage upstream of each base (R^{30} function *filter*, with the *convolution* option, coefficients [1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0]);

4) the ratio of the two convolved signal was calculated for each base, yielding a normalized convolved signal in the interval [-1;1], where, at a given base, positive or negative values indicate a drop or an increase in coverage respectively. In particular, 1 indicates a full drop in coverage from *n* to 0;
5) local maxima of the normalized convolved signal were calculated, indicating genomic locations where the coverage drop is most pronounced. The signal at those locations was called RTS (reverse transcriptase stalling) value.

Identification of RTS sites. For each replicate and for each local maxima (identified as described above):

1) bases with single-base coverage below 6 and coverage drop signal (RTS) below 0.2 (20% of reads stalling) were removed to eliminate low-confidence data points and reduce the number of statistical tests performed in the following step;
2) two conditions were selected, for instance K^+ and Li^+, and the normalized convolved coverage signal for both was used to fit a linear model (function *lm* in R) and estimate the p-value of the fitting through ANOVA testing;
3) all p-values from the linear models contrasting any two conditions were then corrected for multiple hypothesis testing by applying FDR correction on all tested local maxima;
4) significant regions were identified as those having FDR≤0.1 and referred to as scoring regions, or "RTS sites"; for comparison, regions with p-value≤0.01 were scored (**Supplementary Fig. 4b,c**). Sites with RTS value<0.25 and not overlapping any other sites (cases of overlapping G-quadruplex structures or multiple structural isoforms) were further removed from the analysis to further remove regions with subtle effects.

The scale of the RTS value is 0-1. For instance, an RTS value of 0.3 indicates that 30% of the reads containing the rG4 motif have shown stalling, and suggests that the rG4 structure was present in 30% of the population of a particular transcript (or at least, was stable enough to cause stalling in 30% of transcripts).

Li^+ scoring regions anlaysis. The 6,299 Putative Quadruplex sequences (PQs) were analysed: the coverage within the PQ and in the 30 base pairs immediately upstream of the PQ was calculated for each Li^+ replicate independently; coverage values inside and downstream of PQs were normalized (rpkm); the 4 downstream and inside PQ coverage values were fitted to a linear model and the p-value calculated through ANOVA testing; p-values were adjusted for multiple hypothesis by using FDR correction. The sites exhibiting FDR ≤ 0.1 and stalling value ≥ 0.25 were extracted and only the ones not overlapping to K^+ nor K^++PDS scoring regions were reported (**Supplementary Table 5;** *n=72* sites, **Supplementary Note 3**).

G and GG motif analysis after stalling site. We identified the start site of the RTS stalling sites (i.e., the base immediately adjacent to the identified coverage drop) and we analyse whether a G or GG motif is present within 1 nucleotide or 3 nucleotides distance from the stall site. The procedure is repeated independently for K^+ and K^++PDS stalling sites, and the percentage of sites exhibiting each motif is reported.

Overlap of K^+ and K^++PDS scoring regions. The number of K^+ scoring regions overlapping to those scoring in K^++PDS was calculated both for all all rG4s and for the rG4s excluding the "Others" category, by using the command intersectBed from the bedtools and requiring at least 80% sequence overlap (option –f 0.8) to generate the Venn diagrams describing the overlap. For the scatter plot analysis of the RTS value in common scoring regions, we considered only the 2,277 rG4s that were isolated, i.e. not overlapping even one base to any other rG4, in order to have a true reflection of the fraction of stalling reads. Pearson correlation coefficient was calculated for the RTS values in K^+ and K^++PDS in this subset.

Correlation heatmaps. For the overall library comparison (**Supplementary Fig. 3b**), we calculated the coverage at each exon for the 12 different libraries and computed the Pearson correlation coefficient (PCC) of the exon coverage values between all library pairs. For the putative G-quadruplex site comparison (**Supplementary Fig. 4a**), we calculated the coverage for all ~6,000 exonic putative G-quadruplex sites (PQs, canonical motif $G_3L_{1\text{-}7}$) and the coverage 30 nt downstream of each PQ. Then, we filtered out PQs with Li^+ coverage below 6 in any replicate, leaving 2,688 PQs for the analysis, and we calculated the RTS value as:

$$RTS\_value = (coverage_{downstream} - coverage_{PQ}) / coverage_{downstream}$$

Finally, we computed the PCC of the RTS_value between all library pairs.

Assignment to transcripts. Scoring regions were identified at genomic locations displaying a drop in coverage consistent across different replicates, as explained in the previous Methods sub-section. To transform genomic coordinates to transcript coordinates, we performed the following steps:

1) FPKM (Fragments Per Kilobase per Million of mapped reads) were calculated for each transcript isoform in the Li^+ condition using the software cufflinks (http://cole-trapnell-lab.github.io/cufflinks/) and transcripts with FPKM≥0.5 were considered as expressed, resulting in 17,622 transcript isoforms belonging to 12,300 unique genes;
2) genomic scoring regions were mapped to exons belonging to the transcript with highest FPKM among the expressed ones for each given gene. If the genomic coordinates would fall in intronic regions, the second most expressed transcript was assessed and so on;
3) transcript coordinates and sequence for the mapped scoring regions were then calculated accounting for intron skipping.

Controls. Reads from control RNA libraries were trimmed using the trim galore software (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) and aligned using the *bwa mem* software (http://sourceforge.net/projects/bio-bwa/files/) to a customized genome comprising of the 4 control sequences in FASTA format. Bam files containing aligned reads were then processed with the bedtools (http://bedtools.readthedocs.org/) to calculate coverage along each control sequences and extract count for each read starting at a given position. From there, the stalling probability was calculated in *R* as the percentage of reads starting at a given position divided by the total reads aligned to the given control (**Supplementary Fig. 2**).

Hierarchical assignment of rG4s. Sequences extending from the stalling site to 50 base pairs upstream in the transcript were extracted for each RTS site and assigned to different rG4s structural sub-classes (**Fig. 1b,c**), defined as follows (regular expressions used for pattern matching shown in brackets): G_3L_{1-7}, canonical rG4s with loop length between 1-7nt ('$(G_{3+}N_{1-7})\{3,\}G_{3+}$', with N = A, U, C or G); long-loops, rG4s with any loop of length >7nt, up to 12 nt for lateral loops and 21 for the central loop (e.g., "$G_{3+}N_{8-12}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$" or '$G_{3+}N_{1-7}G_{3+}N_{13-21}G_{3+}N_{1-7}G_{3+}$'); bulges, rG4s with a bulge of 1–7nt in one G-tract or multiple 1nt bulges (e.g., '$G_{3+}N_{1-9}G_{3+}N_{1-9}(GGH_{1-7}G|GH_{1-7}GG)N_{1-9}G_{3+}$' or '$(GGHG|GHGG)N_{1-9}$ $(GGHG|GHGG)N_{1-9}G_{3+}N_{1-9}G_{3+}$', with H = A, U or C); 2-quartet, rG4s with 4-tracts of two consecutive Gs ('$(G_{2+}N_{1-9})\{3,\}G_{2+}$'); G≥40%, sequences that contain more than 40% G-content and do not fall into the four previous categories; others: not in any previous category (**Fig. 1c**). When matching multiple categories, a region was assigned to the class with higher predicted stability[12], i.e. (from first to last), canonical rG4s, long loops, bulges, 2-quartet.

MEME motif analysis of the "Others" hits. The 50 base pair sequences from the scoring regions mapping to the category "Others", respectively n = 463 and n = 2,065 for K^+ and K^++PDS, were submitted for motif discovery analysis by the meme-chip software (http://meme-suite.org/tools/meme-chip), using default parameters. The 6 most significant motifs and their E-values are reported.

C-content and C-motif analysis. Only exonic canonical rG4s (category G_3L_{1-7}) with coverage ≥ 20 in the Li^+ condition were considered for this analysis, yielding 1,153 detected versus 1,001 undetected G_3L_{1-7} in K^+ and 1,628 detected versus 790 undetected G_3L_{1-7} in K^++PDS. Each rG4 sequence was equally extended upstream and downstream to enclose a total region of 90 nt including both the rG4 motif and flanking regions. The occurrence of C, CC, CCC, and CCCC motif was counted within each extended region containing detected or undetected rG4s and divided by the region length, yielding a motif density value per sequence. Similarly, the occurrence of U, A and G motif was counted (**Supplementary Fig. 9**). The average value of detected and undetected G_3L_{1-7} density was computed and the ratio of the average detected to averaged undetected density was calculated and plotted (**Fig. 1d**). Values < 1 indicate higher presence of a given motif within the undetected versus the detected rG4s, and progressively lower values indicate higher presence in the undetected G_3L_{1-7}, suggesting the a motif is depleted among the detected G_3L_{1-7}. The analysis was repeated for the K^+ and K^++PDS conditions independently.

Delta free energy analysis. For the same G_3L_{1-7} considered in the C-content and C-motif analysis, G_3L_{1-7} were equally extended upstream and downstream to enclose a total regions of 90nt including both the rG4 motif and flanking regions. The free energy of ensemble of these sequences were calculated using the RNAFold software of the ViennaRNA package[15] (http://www.tbi.univie.ac.at/RNA/) and the detected and undetected G_3L_{1-7} were compared (**Fig. 1e**).

mRNA location, relative enrichment and bin analysis. The overlap of each scoring region (RTS site) with 5'UTR, CDS (coding sequences) and 3'UTR was calculated (*bedtools intersect*; **Fig. 2a**). rG4s partially overlapping multiple regions (e.g., 5'UTR and CDS) were

assigned to both regions and counted twice. The number of overlapping regions to each of the three annotated features was then divided by the total region size in base pairs and multiplied by 1000, therefore yielding the rG4 density per kilobase (density per kb; **Fig. 2b**). The average transcript profile showing preferential distribution of the rG4 (**Fig. 2c**) was calculated as follows: each profiles was normalized to the same length and its 5'UTR, CDS and 3'UTR were divided into 5, 40 and 40 bins respectively, roughly reflecting the relative size of 1:8 (5'UTR to CDS) and 1:1 (3'UTR and CDS) of the three regions for all annotated transcripts in the human transcriptome. Then, for each RTS site, the belonging to a given bin was assessed and 1/bin_size (count normalized by bin size in bp) was computed. Finally, all normalized counts per bin were averaged for scored regions in all transcripts and plotted.

Regulatory sites analysis. MicroRNA target sites and polyadenylation signal (PAS) sites were obtained from the TargetScan database of predicted miRNA target sites[31] (http://targetscan.org) and from the GENCODE project[32] (http://www.gencodegenes.org, release 19) respectively. Genomic coordinates were transformed into transcript coordinates in order to calculate distance by skipping introns. Independently for each regulatory feature class and for each scoring region (RTS site), the distance between the region and the closest feature was calculated, assigning 0 for overlapping features. Then, scoring regions were randomly shuffled by uniform resampling across all the expressed transcripts after merging overlapping exons, in order to avoid over-representation of genes with several alternative isoforms. The cumulative distributions of pairwise feature-region distances were built for RTS sites and random regions (**Fig. 2d,e**). Similarly, the cumulative distribution of distance was assessed separately for rG4s up- and downstream of the respective regulatory sites. The fraction of rG4s in proximity (i.e. ≤100 nt) of regulatory sites was calculated for all rG4s and for up- and downstream rG4s (barplots in **Fig. 2d,e**), and compared to random by using the Chi-squared test for proportions (function *prop.test* in *R*).

RNA structure prediction and PPV comparison. As some scoring regions (RTS sites) could be assigned to multiple overlapping genes, we extracted only unique sequences for the structural folding analysis, yielding 3,358 sequences from the initial 3,383. Further, some rG4s identified were overlapping with each other: if the G-quadruplex motif assigned through the hierarchical motif analysis was coinciding, we removed redundant RTS as they would lead to the same structural prediction. This filtering resulted in 3,319 sequences in K^+ for folding analysis. We extended the 50 nt scoring regions by 100nt up- and downstream, resulting in regions of 250 nt. The RNAFold software (http://www.tbi.univie.ac.at/RNA/) was then used for the structural prediction of the 250 nt sequences, and the prediction was repeated with and without imposing single-strand constraint over the rG4 identified motif. We chose to compare the ΔG of the alternative structures (rather than the rG4 structures) for two main reasons: firstly, we want to assess if there is any alternative secondary structure potentially competing with the identified rG4s; secondly, the thermodynamic parameter for the default mode of structure prediction is more complete than the G-quadruplex mode in RNAfold, which is derived from a limited UV-melting dataset used for training the G-quadruplex folding model[33,34].When more than a motif was identified from the structural characterization analysis, the single-strand constraint was imposed over the G-quadruplex

motif closer to the identified stalling site. After, the RNAstructure package[22] (http://rna.urmc.rochester.edu/RNAstructure.html) was used to convert dot bracket (db) files to connectivity table (ct) files and then perform structure comparison between constrained and unconstrained structural prediction of each scoring region (function *CircleCompare;* **Fig. 3b**), which returned the PPV (positive predictive value) shown in the histogram at **Fig. 3a**. In statistics, PPV represents the proportion of positive predictions that are actually true positives, i.e. in the case of structural comparison, it's the fraction of predicted pairs (without the experimental constraint) that occur in the accepted structure (with the experimental constraint).

Cross-species occurrence of rG4s. We have performed a survey of all 68 non-human eukaryotic species with genomes and annotations deposited in the Ensembl genome database, as accessed through BioMart (http://www.ensembl.org). For each of the 3,383 rG4s, we surveyed the presence of genes in other species that are orthologous to the human genes bearing the corresponding rG4s. For a particular $rG4^{h}$ human (h) sequence and a non-human (*non-h*) species, if an ortholog was present for the $rG4^{h}$-bearing gene, we performed a global sequence alignment of the corresponding human and non-human cDNA sequences. The sequence alignment was done using the utilities provided in the *Biostrings* library for *R*[30], with the global alignment option, and substitution matrix, gap extension and gap penalty parameters matching the default options of blastn aligner designated for relatively similar sequences (2, -3, 5 and 2 for match, mismatch, gap opening and gap extension correspondingly, as defaulted on the popular BLAST server for blastn alignment). After the alignment, the $rG4^{h}$ sequence was identified in the aligned human $cDNA^{h}$, using the rG4-seq revealed coordinate of the 3'-end of $rG4^{h}$, along with +10 downstream and -50 upstream sequence range that may well encapsulate most of the quadruplex sequences (rG4s are on average around 30 nt long). Those coordinates were additionally corrected to account for the possible gaps in $cDNA^{h}$ sequence introduced in the $cDNA^{h}$ vs. $cDNA^{non\text{-}h}$ alignment. Next, the matched segment from the aligned non-human $cDNA^{non\text{-}h}$ sequence was examined, corresponding to the region in human $cDNA^{h}$ that engulfs the particular $rG4^{h}$ sequence. The gaps were removed from that segment (if introduced during the alignment) and, if a nucleic acid sequence was left, the presence of a potential G-quadruplex was assessed using the relaxed $(G_{2+}N_{1\text{-}12})_3G_{2+}$ sequence pattern. The latter pattern was in full compliance with 96.3% (3,258) of experimentally verified rG4 structures, with the remaining 3.7% removed from the further analysis. Through this scheme, we identified the G4 sequences that match with the actual experimentally (rG4-seq) validated $rG4^{h}$ by their positioning inside the orthologs. The outcome of this analysis, after iterating across 3,258 (rG4s) × 69 (species, human inclusive) instances, was stored in a matrix, with 1 if there was a rG4 (green in **Fig. 3c**), 0 if there was an ortholog found but without a G4 sequence at the locus matching the human rG4 (magenta in **Fig. 3c**), and NA (non-assigned) for the cases where even an ortholog was absent (grey in **Fig. 3c**). The matrix was then used to cluster the data entries in both rG4 and species dimensions, producing a heatmap where the G4 presence patterns are clustered (**Fig. 3c, Supplementary Fig. 13**).

Gel-based *In vitro* selective 2' hydroxyl acylation. RNA (5 pmol, 7 µl) was mixed with 10 µl of 2× reaction buffer (final conc. 20 mM Tris, pH 7.5, 0.5 mM $MgCl_2$, 150 mM LiCl or 150 mM KCl). The reaction mixture was heated at 95°C for 1.5 min and cooled at 4°C for 1.5 min, followed by 37°C for 15 min for equilibration. At the beginning of the 37°C incubation, 2 µl of nuclease-free water or 50 µM of PDS was added to the reaction and mix thoroughly. After the 15 min incubation, 1 µl of 1M 2-methylnicotinic acid imidazolide (NAI, final conc. 50 mM), synthesized as previously described[35], was added to the reaction to react for 5 min at 37°C. Anhydrous DMSO was used as control. The reactions was quenched by 5 µl of 2 M DTT (final conc. 400 mM), and further cleaned and concentrated with RNA clean and concentrator (Zymo research). The RNA was redissolved in 5.5 µl nuclease-free water and mixed with 1 µl of 5µM Cy5-labeled transcript-specific reverse primer and 3µL of 5× reverse transcription reaction buffer (final conc. 20 mM Tris, pH 7.5, 4 mM $MgCl_2$, 1 mM DTT, 0.5 mM dNTPs, 150 mM LiCl). The mixture was heated at 75°C for 3 min, followed by 35°C for 5 min for primer annealing. After the 5 min incubation, 0.5 µl of Superscript III (200 U/µl) was added and the reverse transcription was carried out at 50°C for 15 min, followed by treatment of NaOH at 95°C for 10 min. Next, 10 µl of 2× stopping solution was added to the reaction mixture. The cDNAs were size fractionated by 8% denaturing polyacrylamide gel.

Code availability. The computer code and scripts to analyse the data are available in **Supplementary Software**.

26 Kwok, C. K., Ding, Y., Shahid, S., Assmann, S. M. & Bevilacqua, P. C. A stable RNA G-quadruplex within the 5'-UTR of Arabidopsis thaliana ATR mRNA inhibits translation. *Biochem J* **467**, 91-102 (2015).

27 Kwok, C. K. & Balasubramanian, S. Targeted detection of G-quadruplexes in cellular RNAs. *Angew. Chem. Int. Ed.* **54**, 6751-6754 (2015).

28 Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**, R36 (2013).

29 Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

30 Team, R. C. R: A language and environment for statistical computing. Vienna, Austria. http://www.R-project.org (2015).

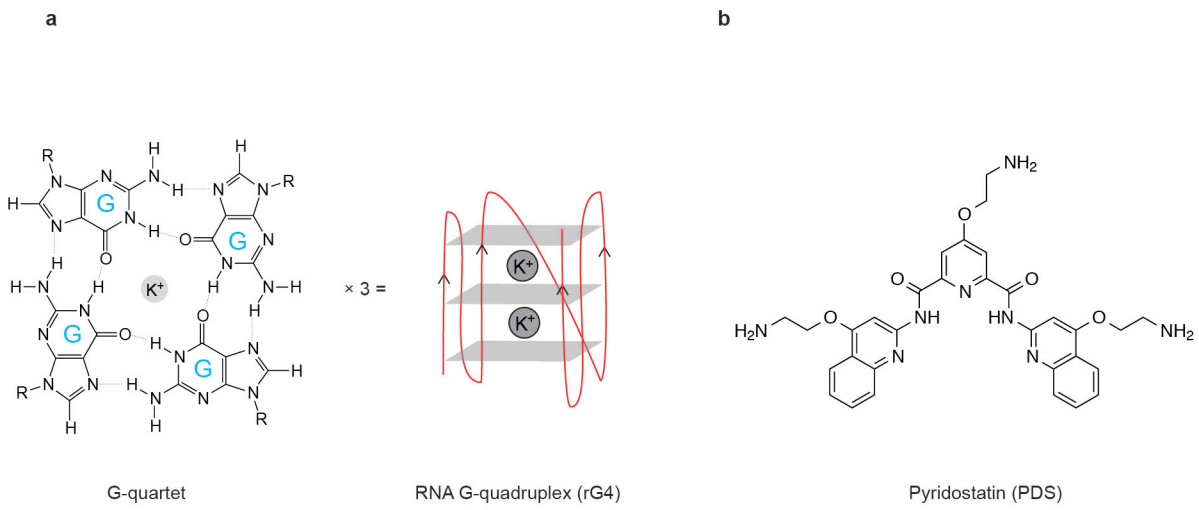
31 Agarwal, V., Bell, G. W., Nam, J. W. & Bartel, D. P. Predicting effective microRNA target sites in mammalian mRNAs. *Elife* **4** (2015).

32 Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**, 1760-1774 (2012).

33 Lorenz, R. *et al.* in *Advances in Bioinformatics and Computational Biology* Vol. 7409 *Lecture Notes in Computer Science* (eds M.C.P. de Souto & M.G. Kann) Ch. 5, 49-60 (Springer Berlin Heidelberg, 2012).

34 Zhang, A. Y. Q., Bugaut, A. & Balasubramanian, S. A sequence-independent analysis of the loop length dependence of intramolecular RNA G-quadruplex stability and topology. *Biochemistry* **50**, 7251-7258 (2011).

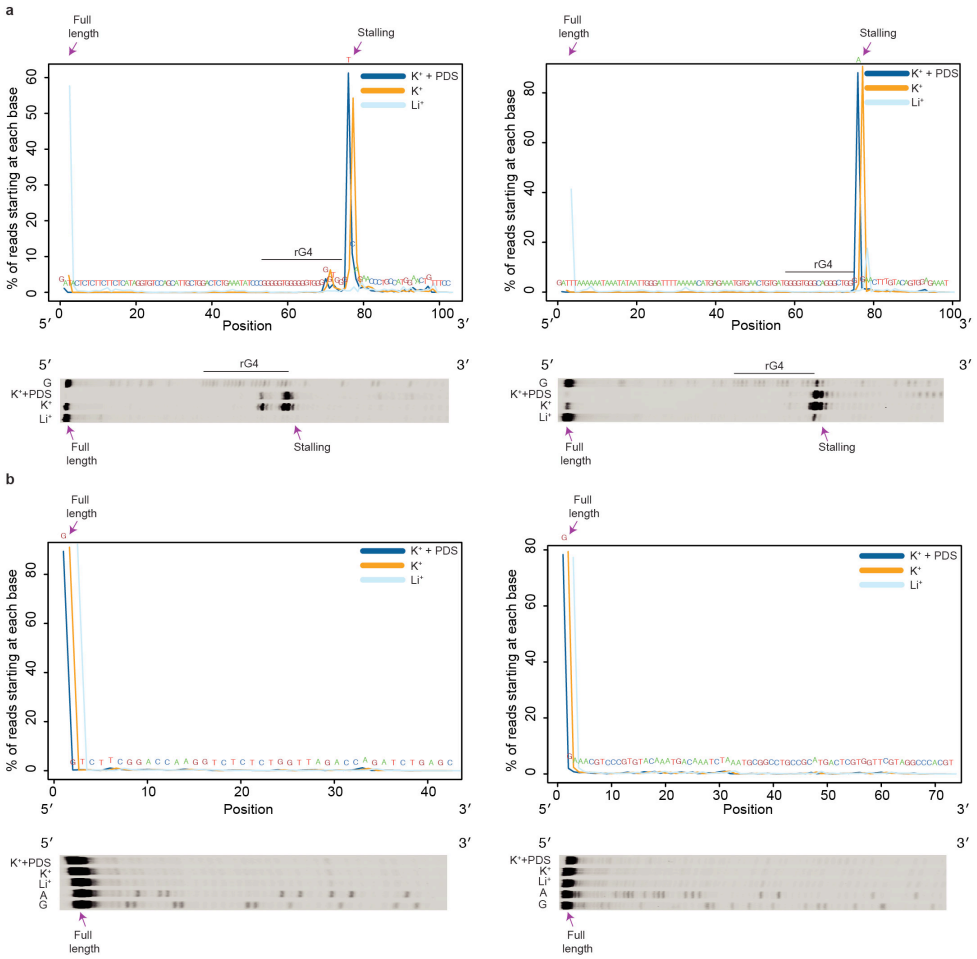
35 Kwok, C. K., Ding, Y., Tang, Y., Assmann, S. M. & Bevilacqua, P. C. Determination of *in vivo* RNA structure in low-abundance transcripts. *Nat Commun* **4**, 2971 (2013).



Supplementary Figure 1

RNA G-quadruplex structure and the stabilizing molecule PDS.

(**a**) Chemical structure of G-quartet and schematic of an intramolecular RNA G-quadruplex (rG4). The presence of K^+ stabilises this RNA structural motif. (**b**) Chemical structure of pyridostatin (PDS), an rG4 stabilising ligand.



Supplementary Figure 2

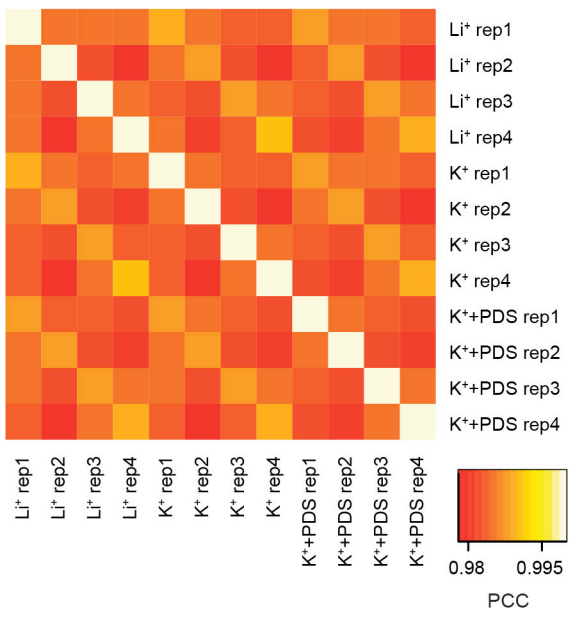
Results from rG4-seq are consistent with the gel-based RTS assay on positive and negative control RNAs.

(**a**) rG4-seq profiles of positive controls (lines), compared to those obtained by gel-based RTS assay (gels) under Li^+, K^+, and K^++PDS conditions. The K^+ and Li^+ rG4-seq data was offset (to the right) by 1 and 2 nucleotide(s) to the K^++PDS data for better visualization. Purple arrows indicate the same base on both line plots and gels. The reverse transcription is from 3' to 5' direction. The rG4-seq results on positive controls (line plots) show strong RTS in K^+ and K^++PDS conditions and no or weak stalling in Li^+ condition, consistent with the corresponding gels (gels). Dideoxy C was used to show G nucleotide in the gels. (**b**) rG4-seq profiles of negative controls (lines), compared to those obtained by gel-based RTS assay (gels) under Li^+, K^+, and K^++PDS conditions. The K^+ and Li^+ rG4-seq data was offset (to the right) by 1 and 2 nucleotide(s) to the K^++PDS data for better visualization. Purple arrows indicate the same base on both line plots and gels. The reverse transcription is from 3' to 5' direction. The rG4-seq results on negative controls (line plots) only show full-length products, and no observable stalling in all three conditions (gels). Dideoxy C and dideoxy T were used to show G and A nucleotides in the gels respectively.

a

Sample	Total reads (M)	Mapped reads (M)	Mapped rate (%)
Li^+ rep1	151.5	125.4	82.8
Li^+ rep2	111.3	93.2	83.8
Li^+ rep3	95.3	83.3	87.4
Li^+ rep4	177.5	142.8	80.4
K^+ rep1	112.9	92.5	81.9
K^+ rep2	93.9	80.5	85.8
K^+ rep3	95.7	80.9	84.5
K^+ rep4	128.5	104.5	81.3
K^++PDS rep1	106.5	82.9	77.8
K^++PDS rep2	101.0	86.2	85.3
K^++PDS rep3	95.7	84.8	88.7
K^++PDS rep4	115.2	92.0	79.8

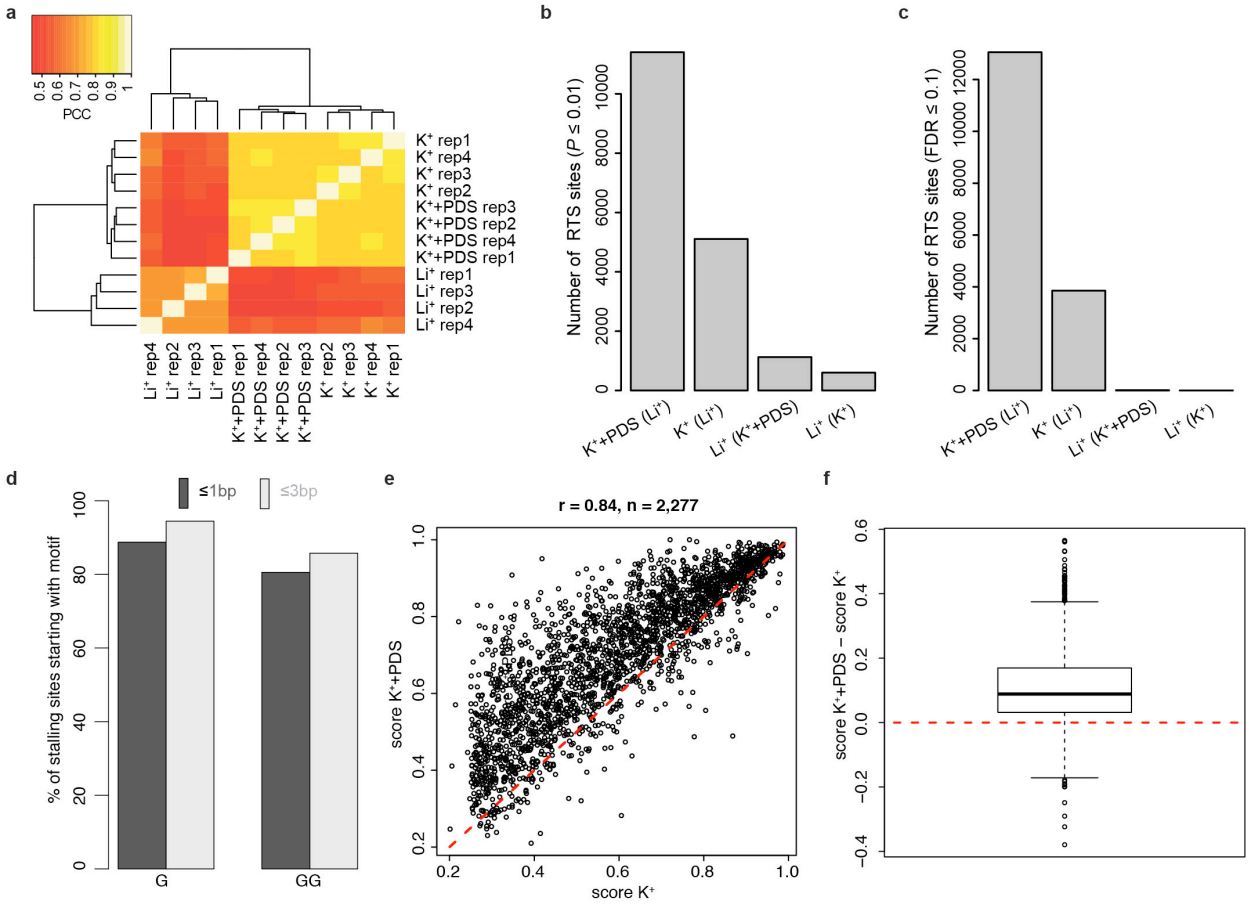
b



Supplementary Figure 3

rG4-seq libraries are highly reproducible.

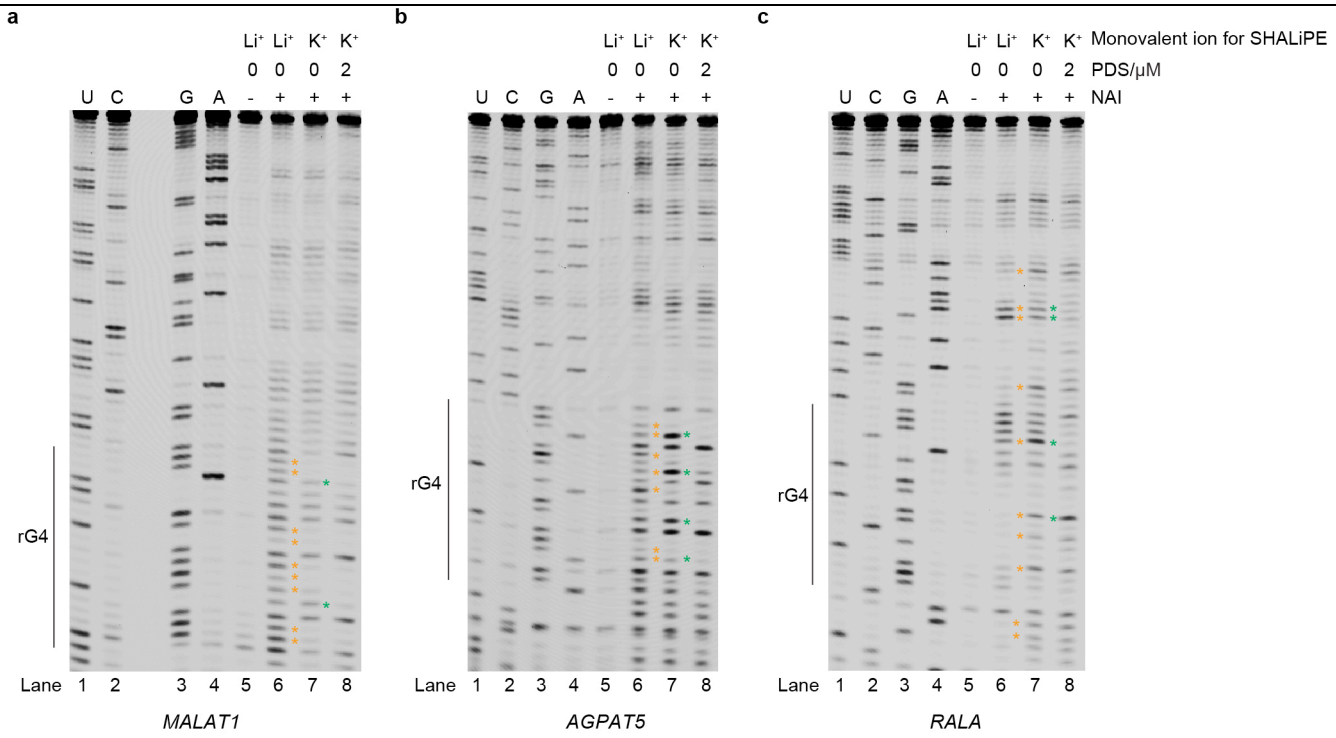
Summary of sequencing libraries and their correlations. (**a**) Table showing total sequenced reads, total mapped reads, and mapping rate for each sample. Four independent biological replicates are performed in this study for each condition (Li^+, K^+, K^++PDS). (**b**) Heatmap showing a color-coded representation of the Pearson correlation coefficient (PCC) of read counts across exons for each pair of libraries, identifying an overall high correlation among the sequencing libraries.



Supplementary Figure 4

rG4-seq scoring pipeline is robust and specific for identification of rG4s transcriptome-wide.

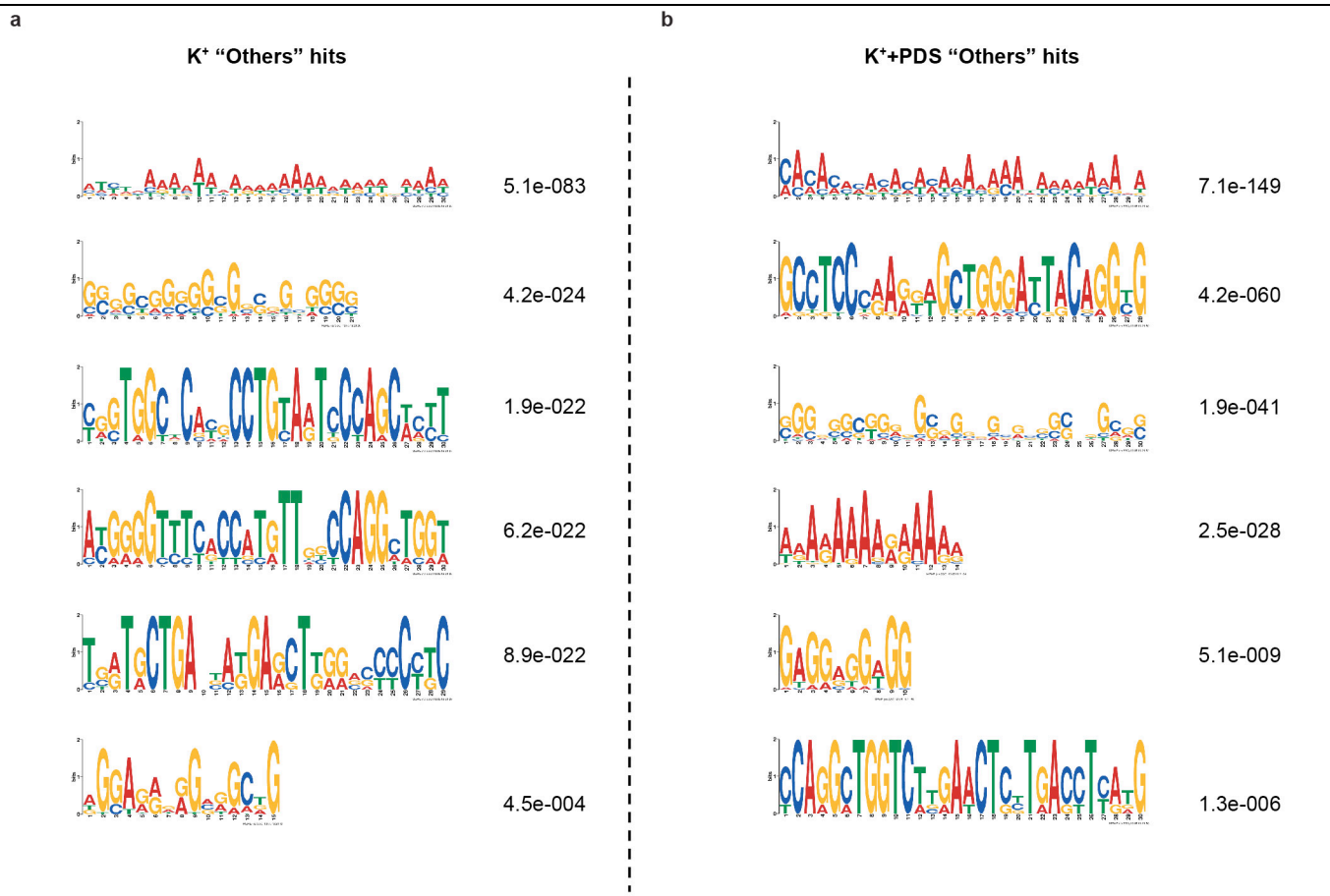
(**a**) Heatmap and hierarchical clustering displaying the similarity of coverage signal for all libraries at the 2,688 exonic putative canonical G-quadruplex sites with coverage ≥ 6 in all Li^+ replicates (see Methods). The color-coded matrix values represented the Pearson correlation coefficient (PCC) for each pair of libraries. As expected, correlation within the same condition (blocks along diagonal) is the highest (0.68-0.84) for all libraries, with Li^+ displaying the lowest within-condition values (0.68-0.70) due to a mild or no effect at rG4 sites. rG4 stabilizing conditions (K^+ and K^++PDS) display instead the highest within-condition correlation (0.81-0.84), and also between the two conditions (0.79-0.82) as they both stabilize similar structures, while showing poor correlation with Li^+ (0.50-0.64 for K^+ and 0.45-0.58 for K^++PDS). (**b**)**,** Barplot displaying the number of scoring regions as assessed by the scoring pipeline (see Methods) when setting the significance threshold to p-value (*P)* ≤ 0.01. (**c**) Same as (**b**)**,** with significance threshold set to FDR (false discovery rate) ≤ 0.1. In both (**b**) and (**c**)**,** the scoring of each region is assessed by comparing the RTS signal in one condition versus another, according to the following legend: K^++PDS (Li^+) = signal in K^++PDS versus Li^+; K^+ (Li^+) = signal in K^+ versus Li^+; Li^+ (K^++PDS) = signal in Li^+ versus K^++PDS; Li^+ (K^+) = signal in Li^+ versus K^+. The FDR-based scoring yields very few regions where RTS is affected in Li^+ (most right bars, values of 13 and 4), which are clearly false positive, while returning many regions where RTS is specific for the two rG4 stabilizing conditions, K^+ and K^++PDS. (**d**) G and GG motifs at detected stalling events. The bar graphs show the percentage of sequences displaying a G (left bars) or a GG (right bars) motif at stalling sites, either within 1 nucleotide (dark grey) or 3 nucleotides (light grey) from the detected stalling sites. The high percentage of G and GG motifs, typical of G-quadruplex forming structures, suggests that stalling events occur exactly at or near G-quadruplex sites. N = 3,845 from the K^+ RTS sites. (**e**) RTS values at stalling sites common between K^+ and K^++PDS. Scatter plot comparing the fraction of stalling reads (RTS value) for the hits common in K^+ and K^++PDS (see Methods). (**f**) Boxplot showing the difference of the RTS values in K^++PDS and K^+ for all the data points shown in (**e**).



Supplementary Figure 5

In vitro selective 2'-hydroxyl acylation experiments validate non-canonical rG4 candidates identified by rG4-seq.

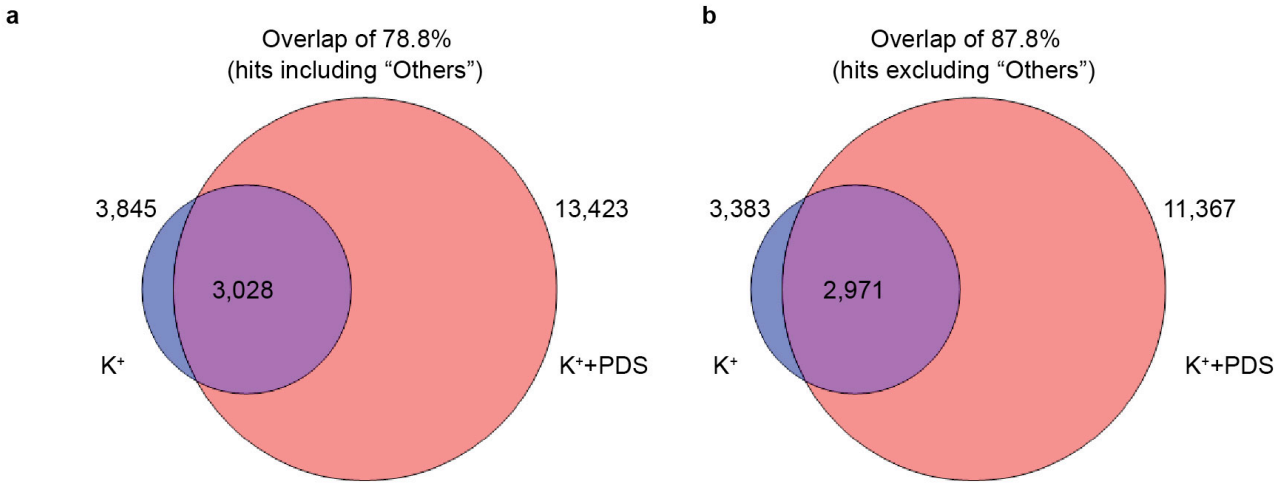
(**a**) *MALAT 1* (chr11:65,269,314-65,269,406). (**b**) *AGPAT5* (chr8:6,617,768-6,617,857). (**c**) *RALA* (chr7:39,726,284-39,726,373). Lanes 1-4 show sequencing of U, C, G, and A respectively. Lane 5 shows the minus 2-methylnicotinic acid imidazolide (NAI). Lanes 6-8 shows the NAI reaction under Li^+, K^+, and K^++PDS conditions respectively. The change in NAI modification suggests change in RNA structural conformation. G-quadruplex structure is stabilized in K^+ and K^++PDS, but not in Li^+, thus the modification change observed was likely attributed to rG4 formation. Orange asterisks denote nucleotides that are changed between Li^+ and K^+ condition. Green asterisks denote nucleotides that are changed between K+ and K^++PDS conditions. Sequences used are shown in **Supplementary Table 1**.



Supplementary Figure 6

MEME motif analysis of the hits in the "Others" category.

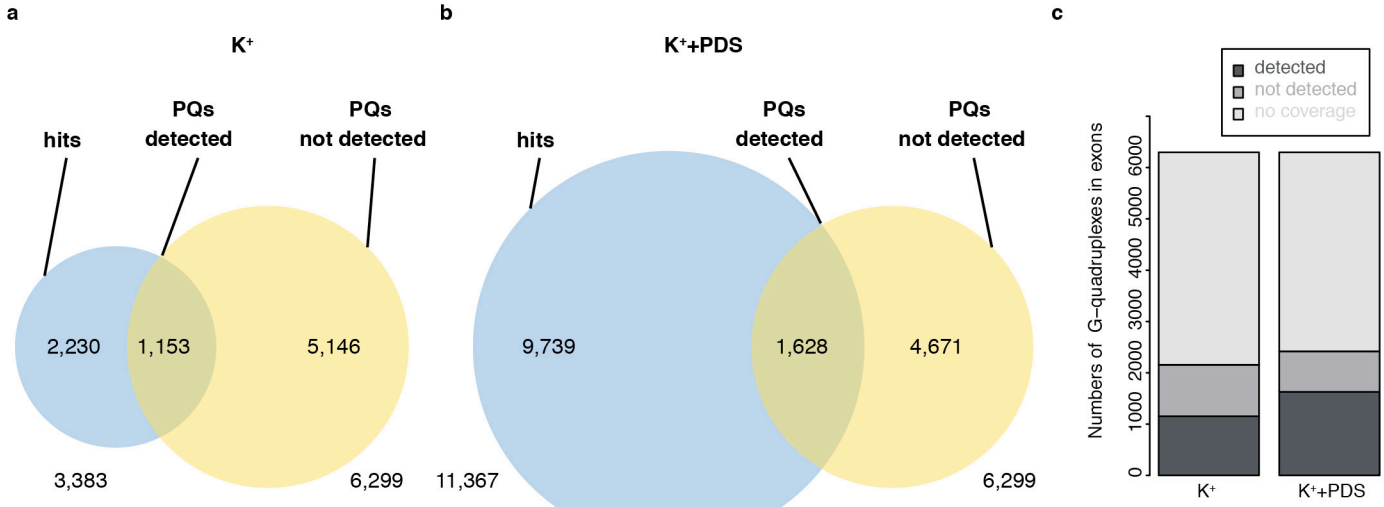
Top 6 motifs as identified by MEME enrichment analysis for the hits in the category "Others" in (**a**) K^+ and (**b**) K^++PDS. Most enriched motifs are non-G-rich, although G-rich motifs are also identified as enriched (2^{nd} and 6^{th} motif in K^+; 3^{rd} and 5^{th} motif in K^++PDS).



Supplementary Figure 7

Overlap of rG4s in K^+ and K^++PDS.

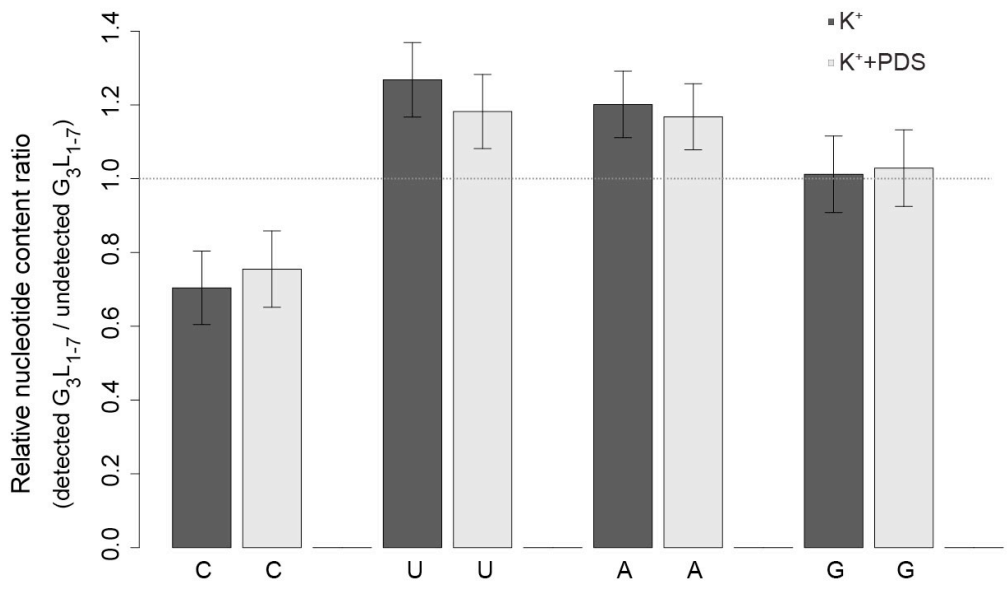
(**a**) Venn diagram showing the overlap for all hits, including those in the “Others” category. (**b**) Overlap for hits, excluding those in the “Others” category.



Supplementary Figure 8

Overlap between hits and computationally predicted G-quadruplex structures (PQs).

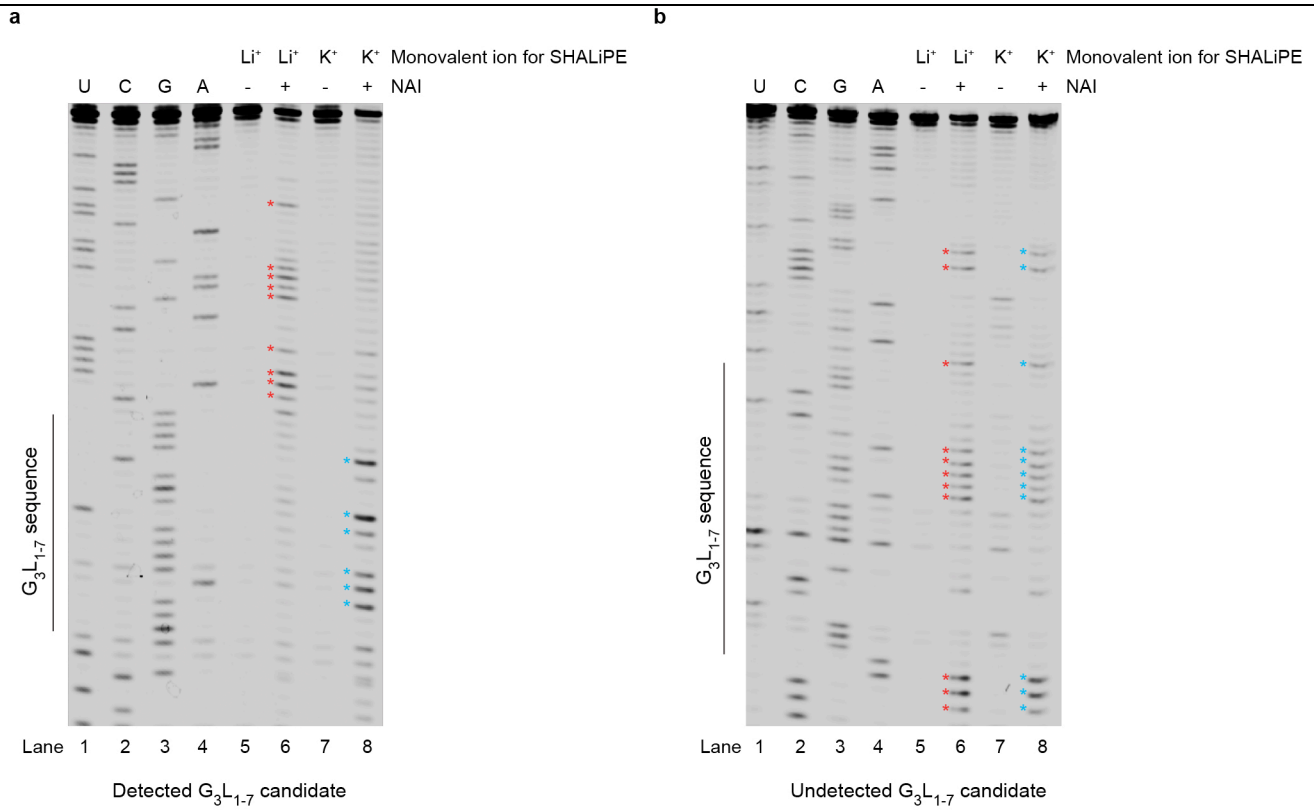
(**a**) Venn diagram showing the overlap between scoring regions in K^+ and PQs ($G_{3+}L_{1-7}$, see Methods). (**b**) Same as (**a**) for scoring regions in K^++PDS. (**c**) Bar plot showing the repartition of the 6,299 exonic PQs between detected as hits (label “detected”, dark grey), not detected as hit but with sufficient coverage (i.e., coverage above 6; label “not detected”, grey) and with no coverage (i.e., coverage below 6; label “no coverage”, light grey) for both K^+ and K^++PDS conditions.



Supplementary Figure 9

Relative nucleotide content ratio on detected G_3L_{1-7} versus undetected G_3L_{1-7} for K^+ and K^++PDS.

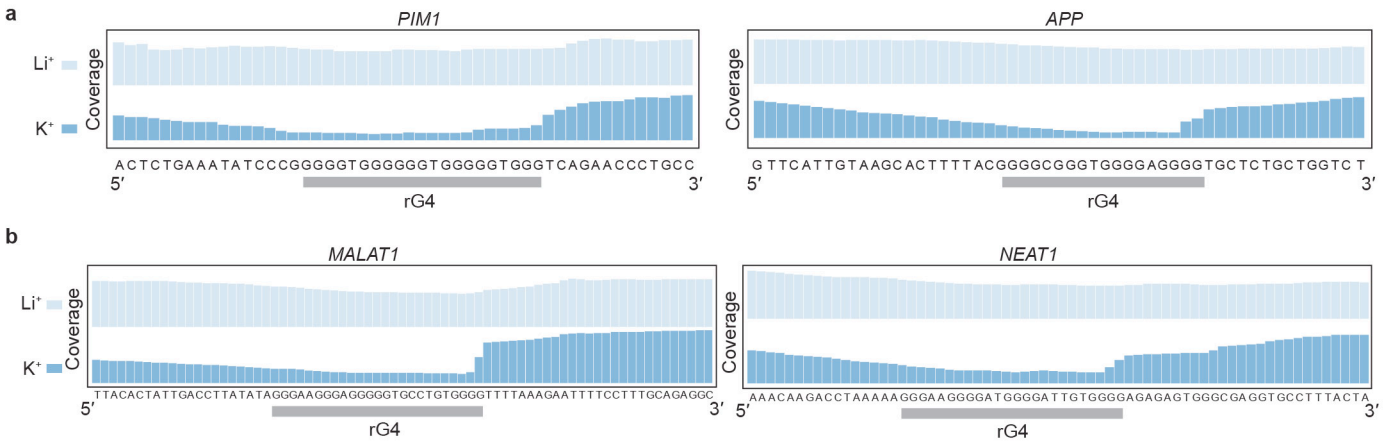
The detected G_3L_{1-7} have a lower C-content than undetected G_3L_{1-7} case for K^+ and K^++PDS (smaller than one), while the detected G_3L_{1-7} have a higher U-content/A-content than undetected G_3L_{1-7} case for K^+ and K^++PDS (larger than one). Errors are calculated from standard deviation. The two bars on C are identical to **Fig. 1d** (single C-motif), and are shown again here for the purpose of comparison to other three nucleotides.



Supplementary Figure 10

In vitro selective 2'-hydroyxl acylation experiments show change in modification profiles between Li^+ and K^+ conditions on detected G_3L_{1-7} candidate but not for undetected G_3L_{1-7} candidate.

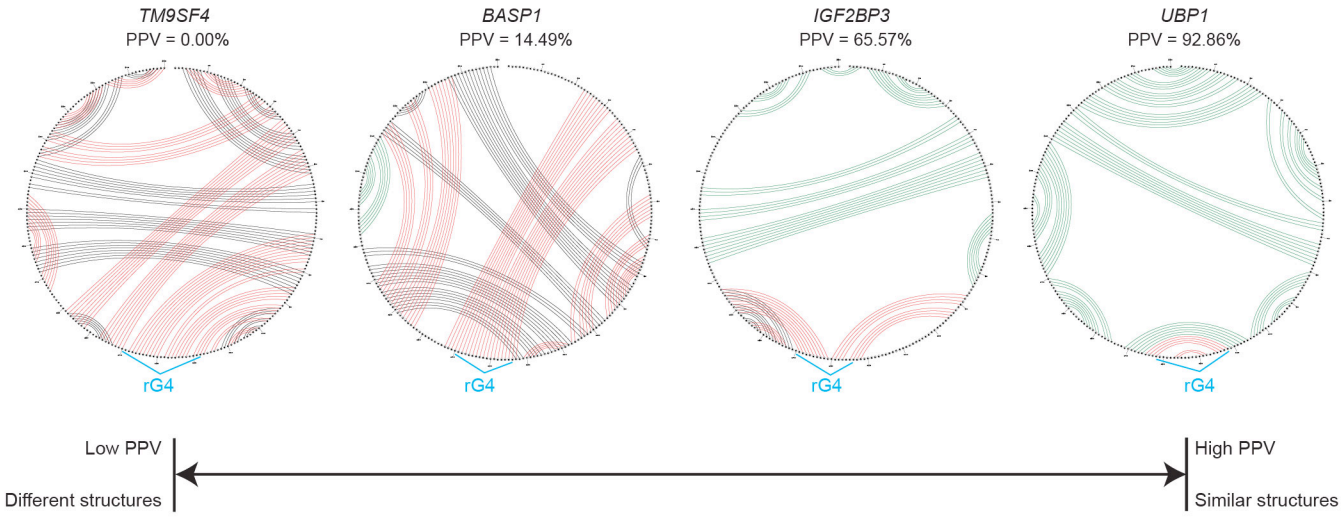
In vitro selective 2'-hydroyxl acylation experiments on examples of (**a**) detected and (**b**) undetected G_3L_{1-7} candidate. For the detected G_3L_{1-7} candidate, *APP* (chr21:27,253,214-27,253,291), the modification profiles for Li^+ (red asterisks) and K^+ (blue asterisks) are strikingly different, i.e. RNA structures are different under Li^+ and K^+ conditions. G-quadruplex structure is stabilized in K^+, but not in Li^+. In contrast, the modification profiles for undetected G_3L_{1-7} candidate, *COMTD1* (chr10:76,993,759-76,993,842), are nearly identical (compare red and blue asterisks), indicate no change in RNA structure. Sequences used are shown in **Supplementary Table 1**.



Supplementary Figure 11

rG4-seq identifies rG4s in mRNAs and lncRNAs.

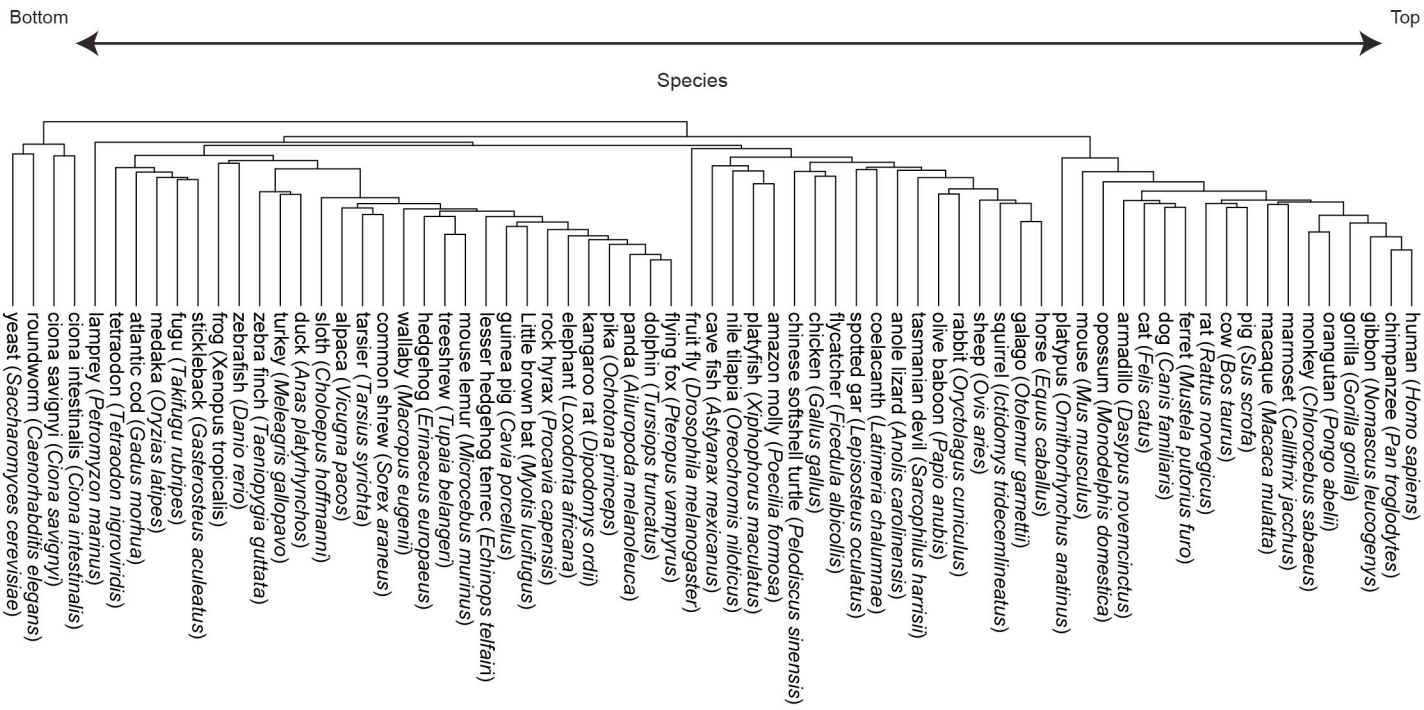
Representative mRNA transcripts harbouring rG4 in (**a**) *PIM1* (chr6:37,142,130-37,142,179) and *APP* (chr21:27,253,230-27,253,282). Representative lncRNA transcripts harbouring rG4s in (**b**) *MALAT1* (chr11:65,271,535-65,271,607) and *NEAT1* (chr11:65,193,478-65,193,543).



Supplementary Figure 12

Comparison of RNA secondary structures with or without rG4 constraints uncovers local to global change in RNA conformation.

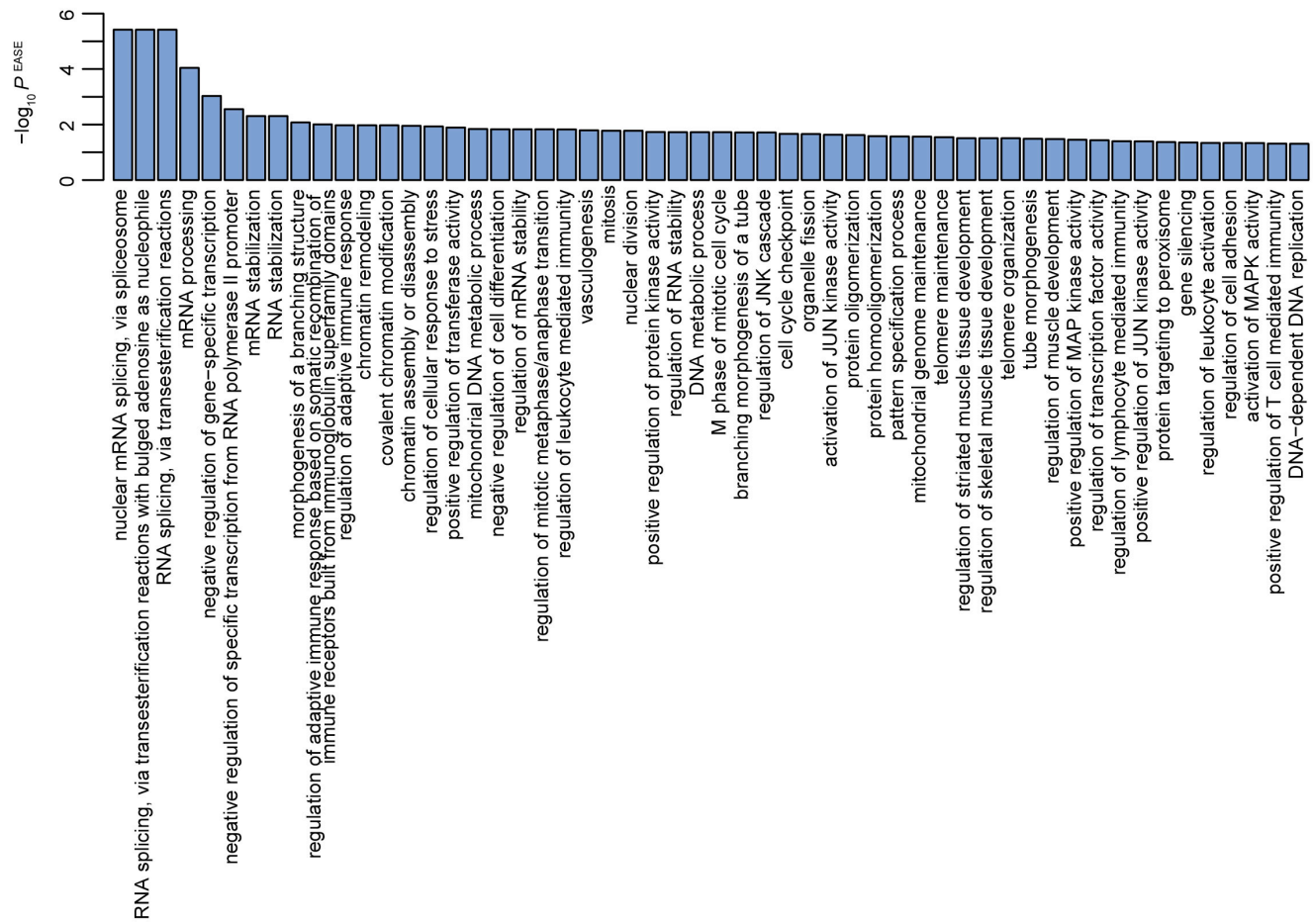
Representative examples of RNA secondary structures with or without rG4 constraint. *TM9SF4* (chr20:30,753,266-30,753,515), *BASP1* (chr5:17,276,089-17,276,338), *IGF2BP3* (chr7:23,351,620-23,351,869), *UBP1* (chr3:33,481,621-33,481,870). The structural comparisons are analysed using RNAstructure[1] and visualized using CircleCompare. Green, base pairs present in both structures. Red, base pairs present only in structure without rG4-constraint. Black, base pairs present only in structure with rG4-constraint. Low PPV indicates different in structures.



Supplementary Figure 13

Common and scientific names of the analysed species.

The clustering of the species corresponds to the one shown in **Fig. 3c**. Taking into account the presence of many non-assigned entries, where orthologs were absent (grey in **Fig. 3c**), binary distance metric was used for clustering, as implemented in the heatmap.2 function of the gplots library for R^2. Expectedly, the species that are clustered close to the human, based on the presence of analogous rG4s, are the hominoid apes (**Fig. 3c**). The image has been rotated 90° clockwise for visualization purposes.



Supplementary Figure 14

Significantly enriched GO terms that are exclusive to strong CSO group.

The 54 GO terms corresponding to **Fig. 3e** are shown. The data come from gene ontology (GO)[3] term (BP set) enrichment analysis for all the unique genes from each of the strong and average CSO groups outlined in **Fig. 3e**. We used DAVID gene functional annotation server[4] and the frequency of the genes in *Homo sapiens*, as a normalisation background (**Supplementary Table 4**). The terms were declared as significantly enriched with the genes, if possessing a corrected p-value (EASE score)[4] of less than 0.05 (or $-\log_{10}P^{EASE}$ >1.301). The gene group with strong CSO was enriched in 117 GO terms, as compared to 298 terms enriched with average CSO (**Fig. 3d**). The terms here show the ones unique to the strong CSO term.

Supplementary Note 1. Current methods for rG4 identification.

Previous methods to identify rG4 structures in transcripts have employed computational predictions rather than experimental evidence, and the prediction algorithms generally do not consider the effect of flanking sequences and their sequence context[5,6]. Current state-of-the-art for experimental rG4s identification is to use methods such as circular dichroism, UV-melting, nuclear magnetic resonance, and in line probing, on synthetic oligonucleotides one at a time[7-10]. Although they are useful for the identification of rG4s, these low-throughput methods are time-consuming and not practical for the experimental evaluation of rG4s transcriptome-wide.

Supplementary Note 2. Limitation of rG4-seq applied to date.

Like existing low-throughput rG4 identification methods, one limitation of rG4-seq is that it maps rG4 structure *in vitro*, thus the rG4s reported here may differ from the *in vivo* condition. Recent methodologies allow inferring RNA structural properties *in vivo*[11,12]. However, the inspection of recently published dataset with icSHAPE genome-wide data[13] showed that most rG4 regions do not report structural information at these sites because of insufficient coverage. Further development of the SHAPE method, such as utilizing SHALiPE[14], or alternative strategies will be required to enable mapping rG4 structure in living cells.

Supplementary Note 3. Potential drawback of the described RTS scoring method.

One potential drawback of our RTS scoring method is that if a sequence displays a large coverage drop already in Li^+ condition, it could be a potential false negative. To evaluate the extent of this issue, we analysed the coverage drop at canonical rG4s (G_3L_{1-7}) (see Methods). We identified 72 sites with FDR ≤ 0.1 (**Supplementary Table 5**). However, we can not trust these sites as *bona fide* regions of RT stalling due to the presence of a G-quadruplex structure: the coverage drop could be due to a more general (i.e. rG4-independent) lack of sequencing coverage, which is known to happen in sequences of low complexity and high GC richness. We therefore did not include the 72 sites in our list of rG4 scoring regions.

Supplementary References:

1 Reuter, J. S. & Mathews, D. H. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* **11**, 129 (2010).
2 R Core Team: A language and environment for statistical computing. Vienna, Austria. http://www.R-project.org (2015).
3 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29 (2000).
4 Huang da, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57 (2009).
5 Kikin, O., D'Antonio, L. & Bagga, P. S. QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic Acids Res.* **34**, W676-682 (2006).
6 Huppert, J. L., Bugaut, A., Kumari, S. & Balasubramanian, S. G-quadruplexes: the beginning and end of UTRs. *Nucleic Acids Res.* **36**, 6260-6268 (2008).
7 Vorlickova, M. *et al.* Circular dichroism and guanine quadruplexes. *Methods* **57**, 64-75 (2012).
8 Mergny, J. L., Phan, A. T. & Lacroix, L. Following G-quartet formation by UV-spectroscopy. *FEBS Lett.* **435**, 74-78 (1998).
9 Webba da Silva, M. NMR methods for studying quadruplex nucleic acids. *Methods* **43**, 264-277 (2007).
10 Beaudoin, J. D., Jodoin, R. & Perreault, J. P. In-line probing of RNA G-quadruplexes. *Methods* **64**, 79-87 (2013).
11 Kwok, C. K., Tang, Y., Assmann, S. M. & Bevilacqua, P. C. The RNA structurome: transcriptome-wide structure probing with next-generation sequencing. *Trends Biochem. Sci.* **40**, 221-232 (2015).
12 Lu, Z. & Chang, H. Y. Decoding the RNA structurome. *Curr Opin Struct Biol* **36**, 142-148 (2016).
13 Lu, Z. *et al.* RNA duplex map in living cells reveals higher-order transcriptome structure. *Cell* **165**, 1267-1279 (2016).
14 Kwok, C. K., Sahakyan, A. B. & Balasubramanian, S. Structural analysis using SHALiPE to reveal RNA G-quadruplex formation in human precursor microRNA. *Angew. Chem. Int. Ed.* DOI: 10.1002/anie.201603562 (2016).