Supplementary Materials for

RNA G-quadruplexes are globally unfolded in eukaryotic cells and depleted in bacteria

Junjie U. Guo and David P. Bartel

Correspondence to: dbartel@wi.mit.edu

This PDF file includes:

Materials and Methods
Figs. S1 to S9
Reference (*39*−*41*)
Captions for tables S1 to S6

Other Supplementary Materials for this manuscript includes the following:

Tables S1 to S6

Supplemental Materials and Methods

Mammalian cell culture and transfection

Feeder-free mESCs were cultured on 0.1% gelatin-coated plates in DMEM (high glucose) supplemented with 1X nonessential amino acids (Invitrogen), 1X penicillin-streptomycin-glutamine (Invitrogen), 15% defined fetal bovine serum (HyClone-ES screened) and 1000 U/mL ESGRO leukemia inhibitory factor (Millipore). Adherent HeLa and HEK293T cells were cultured in DMEM (high glucose) supplemented with 10% fetal bovine serum. After splitting and culturing overnight to 70−90% confluence, 10 cm plates of HEK293T cells were each transfected with 10 µg pcDNA3-mCherry-G3A2 plasmid and 60 µl Lipofectamine 2000 (Invitrogen) according to the manufacturer's instructions.

Yeast culture and induction

S. cerevisiae strain yRH101 harboring the pYES2.1-mCherry-4xAAGGG plasmid was cultured at 30°C in SD-Ura media supplemented with 2% raffinose. For induced expression, an overnight culture was inoculated into SD-Ura media with 2% galactose at an OD_{600} of 0.4 and allowed to grow for another 6 hours before harvest.

Bacterial culture and induction

For growth curves, overnight cultures of *E. coli* TOP10 strains harboring pCR2.1-mCherry plasmids with RG4 or mutant insertions were diluted to an OD_{600} of 0.1 with LB supplemented with 100 µg/ml ampicillin and allowed to grow at 37°C with shaking. For induced protein expression, exponential-phase (OD_{600} = 0.4−0.6) cultures of BL21 strains that overexpressed LacIq and harbored the plasmids were induced by addition of 0.5 mM IPTG and allowed to grow at 37°C for two hours before harvest. Frozen stock of *P. putida* strain KT2440 was obtained from ATCC (#47054). A single colony was picked from an LB/agar plate and cultured in LB broth at 37°C with shaking. A 400mL axenic culture of *Synechococcus* sp WH8102 was grown in natural seawater-based Pro99 media containing 0.2 µm filtered Sargasso Sea water amended with Pro99 nutrients, supplemented with 6 mM sterile sodium bicarbonate to support the growth of a large volume of culture (*39*). Cells were grown at 24 °C under constant illumination (45 µmol photons m^{-2} s^{-1}) to mid-exponential growth phase.

RNA purification

Total RNA was extracted from mammalian cells using TRIzol (Invitrogen) according to the manufacturer's instructions. Yeast total RNA was extracted with hot acid phenol:chloroform:isoamyl alcohol (PCA, 25:24:1, Ambion) with 0.5% SDS and then ethanol precipitated. Poly(A)$^+$ RNA was purified using oligo(dT) Dynabeads (Invitrogen) according to the manufacturer's instructions. To purify bacterial total RNA, cells were first treated with 2 mg/ml lysozyme (Sigma) for 2 minutes at room temperature before hot PCA/SDS extraction and precipitation. Ribosomal RNA was depleted using Ribo-Zero for Gram-negative bacteria (Epicenter).

Mapping of sequencing reads

The six random nucleotides of the RT primer reduced circLigase bias and enabled the removal of PCR-duplicated sequences. After removing these duplicates, the first six nucleotides and the 3′-adapter sequences were trimmed using FASTX tools. The trimmed sequences were aligned to either the *S. cerevisiae* transcriptome (Saccharomyces Genome Database, version 2015-1-13), non-redundant RefSeq-based mouse or human transcriptomes (available at bartellab.wi.mit.edu/publication.html), or RefSeq transcriptomes for *E. coli*, *P. putida* and *Synechococcus sp WH8102* using Bowtie, requiring unique mapping and allowing one mismatch.

Transcript-specific analyses using primer-extension assays

1 µl 0.2 M Tris-Cl (pH 7.5), 1 µl 1.5 M KCl or NaCl (for DMS probing and NAI probing, respectively), 0.5 µl 60 mM MgCl2, 0.5 µl 10 mM dNTP mix and 0.5 µl 1 µM 5′-^{32}P or ^{33}P-labeled primer ($^{32/33}$p-CAAGCAGAAGACGGCATACG, IDT) were added to 5 µl RNA template (2 µg or 5µg DMS-treated or NAI-treated poly(A)$^{+}$ RNA, respectively). The mixture was incubated at 80°C for 2 minutes then cooled to 42°C and incubated for additional 2 minutes before adding 50 U SuperScript III reverse transcriptase (Invitrogen). After incubation at 42°C for 10 minutes, the reaction was stopped with addition of 1 µl 1 M NaOH, and the mixture was heated at 98°C for 15 minutes to hydrolyze the RNA. The mixture was neutralized with 1 µl 1M HCl, mixed with an equal volume of Gel Loading Buffer II (Ambion) and denatured at 85°C for 15 minutes. cDNAs were separated on a 8% urea gel. ^{32}P gels were frozen and imaged using a Typoon Phosphoimager (GE). ^{33}P gels were fixed in 10% acetic acid and 10% methanol and dried at 80°C under vacuum (Bio-Rad gel dryer) before imaging.

RG4 expression constructs

The G3A2 quadruplex (underlined) was appended by PCR to the *mCherry* CDS (in bold) within the context of the following sequence:
…**TACAAG**TAAATAGATTTGCGTTACTGTCTAAGGGAAGgGAAGgG
AAGGGTTTTTCTTTTATTTTCTTTTCGTATGCCGTCTTCTGCT*TGA*AAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAA. This PCR product was inserted to expression vectors (pCR2.1 for *E. coli*, pYES2.1 for yeast, pcDNA3.1 for mammalian cells) by TOPO TA cloning (Life Technologies). The A_{30} region enabled the enrichment of the transcript from total *E. coli* RNA using oligo(dT) beads. In G3U constructs, the underlined segment was replaced with TTTGgGTGgGTGGGTGGG. For stop-codon mutants, the TAA stop codon was replaced with AAA so that the CDS was extended to include the RG4 region before terminating at the downstream TGA codon (italicized). In G3A2 and G3U mutants, the middle G residues (small letters) in two of the four G tracts were mutated to A and U, respectively.

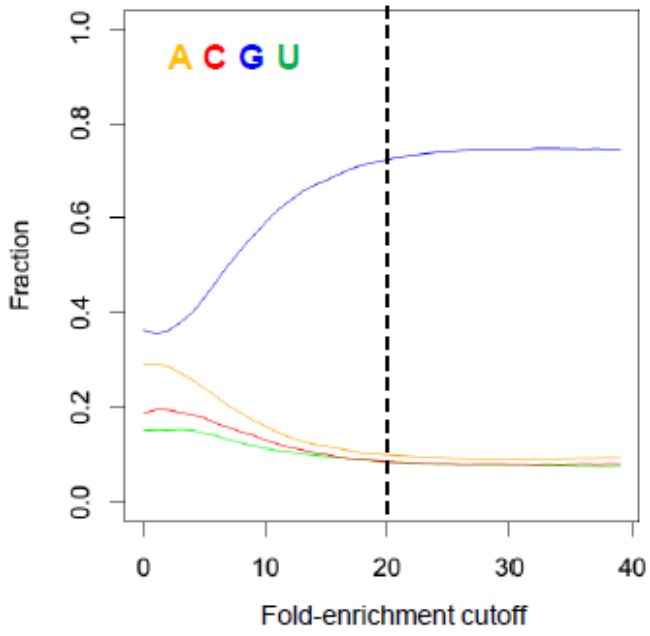

Fig. S1. Nucleotide composition at position 0 of RT stops of varying strength. At each fold-enrichment cutoff, the fraction of stops at the indicated nucleotide is plotted, considering only stops supported by ≥10 reads. Stops with enrichment of ≥20 fold (dashed line) were carried forward as strong RT stops.

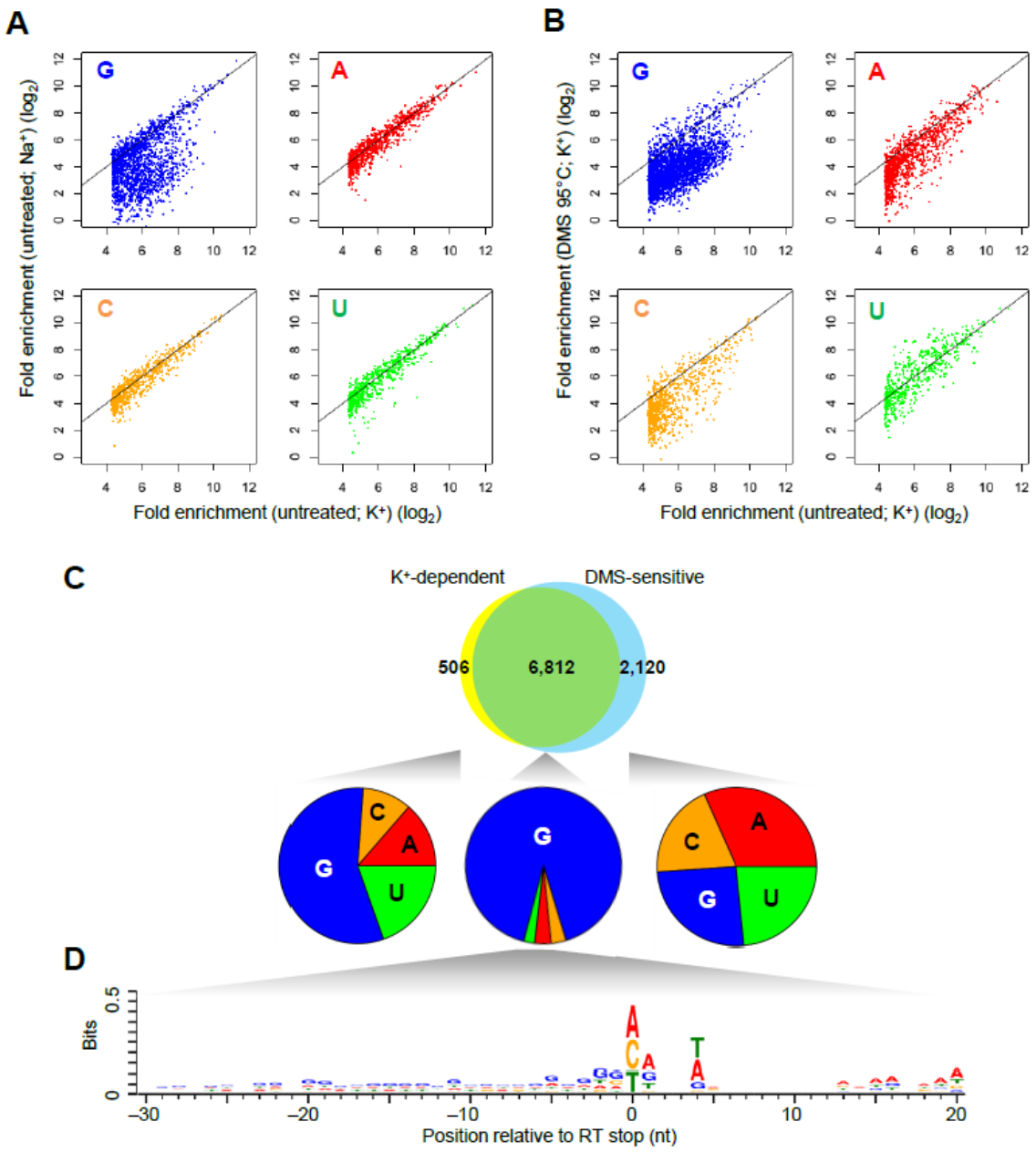

Fig. S2. K^+-dependency and DMS-sensitivity of strong RT stops. (**A**) The results of the top panel of Fig. 2B, re-plotted to show the data for RT stops at each of the four nucleotides separately. (**B**) The results of the bottom panel of Fig. 2B. (**C**) Nucleotide composition at position 0 for RT stops that were either K^+-dependent but DMS-insensitive (*left*), K^+-dependent and DMS-sensitive (*middle*), or K^+-independent but DMS-sensitive (*right*). (**D**) Nucleotide composition of the flanking sequences of 672 K^+-dependent, DMS-sensitive strong RT stops that lacked a G at position 0. Nucleotides are plotted at heights indicating the information content of their enrichment (bits).

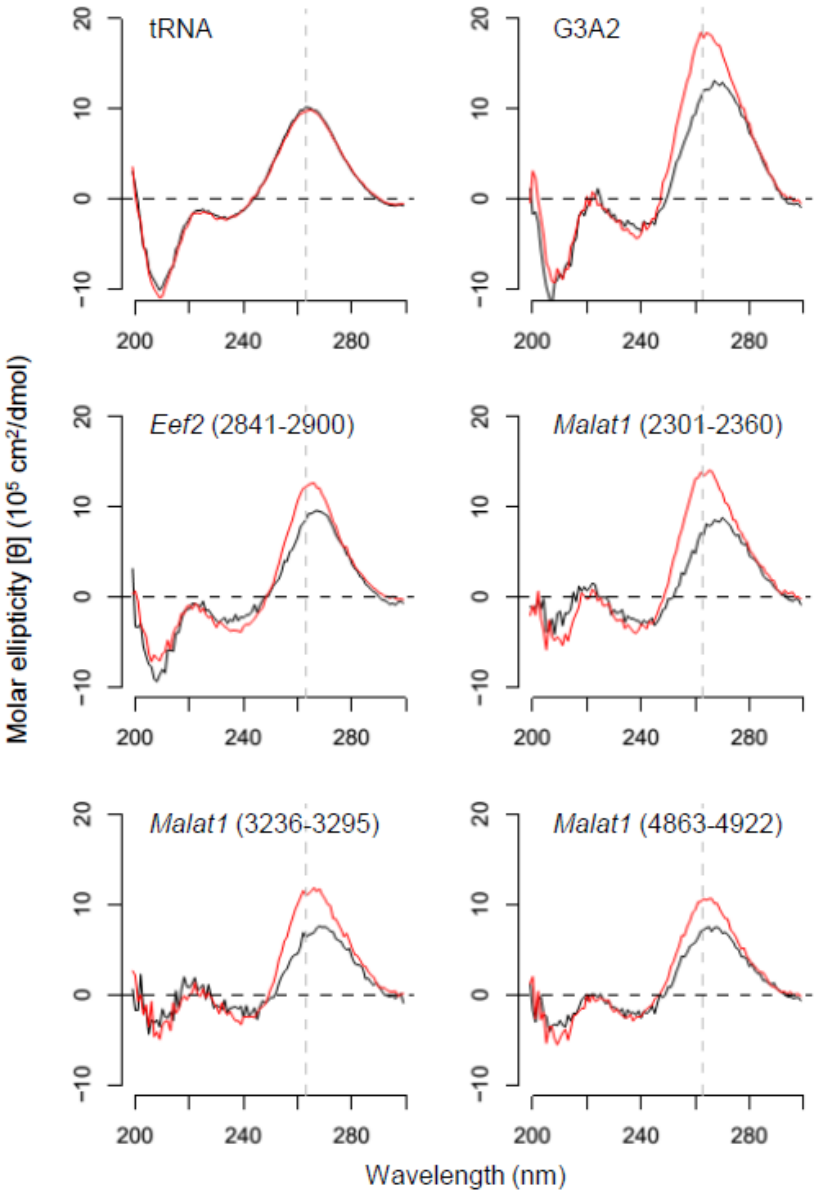

Fig. S3. Circular dichroism (CD) spectra of control RNAs (*E. coli* tRNA and the G3A2 quadruplex) and four 60-nt regions upstream of K^+-dependent strong RT stops. Each RNA (4 μM) was heat-denatured and allowed to fold at 42°C in 20 mM Tris-Cl (pH 7.5), 1mM $MgCl_2$ and either 150 mM (red) or 0 mM K^+ (black). A K^+-dependent increase in the positive peak at 263 nm (vertical dashed line), indicative of parallel RG4 structure formation, was observed in the G3A2 positive control and all four RG4 regions examined, but not in the tRNA negative control. The G3A2 RNA had the sequence GG ATAGATTTGCGTTACTGTCTAAGGGAAGGGAAGGGAAGGGTTTTTCTTTTATT TTCTTTTCGTATGCCGTCTTCTGCTTG. The endogenous RG4 regions included the segments of *Eef2* and *Malat1* transcripts (NM_007907 and NR_002847, respectively) shown in parentheses. To facilitate efficient in vitro transcription, G residues were added to regions that did not already begin with GG at their 5′ termini.

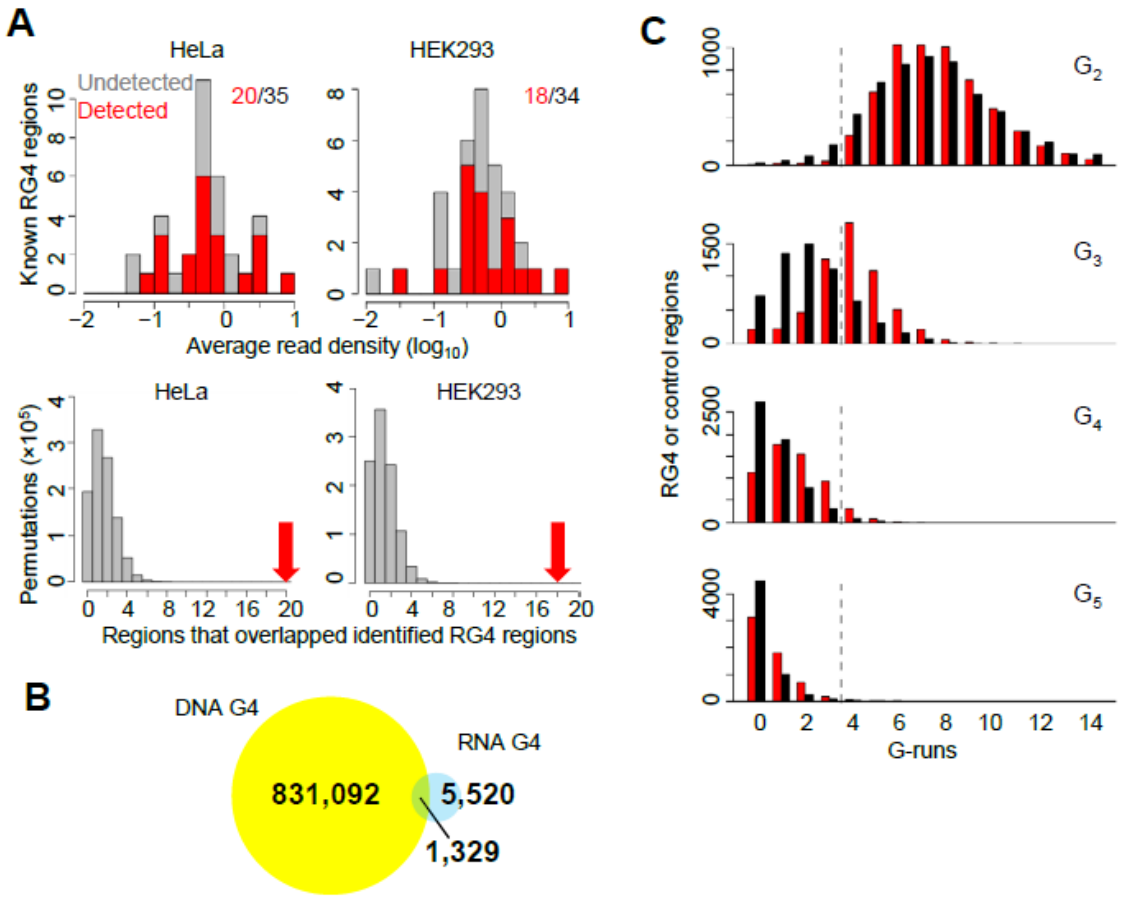

Fig. S4 . Characteristics of RG4 regions. (**A**) Overlap between the previously identified RG4 regions and those identified as K^+-dependent RT stops. The upper panels plot the distribution of the average RT stop read densities of mRNAs with previously known RG4 regions in HeLa cells (*left*) and HEK293T cells (*right*), indicating in red those that overlapped the RG4 regions identified through RT-stop profiling (table S2). All mRNAs with at least one RT-stop read were considered. The lower panels show results of permutation tests using one million cohorts of length-matched randomly selected regions from mRNAs with at least one RT-stop read, showing chance overlap with the previously known regions. The numbers of previously known RG4 regions detected in the upper panels, indicated by red arrows, exceeded those of all permutations. The mean of all permutations, 1.7 and 1.3, respectively, implied false-discovery rates of 1.7/20 and 1.3/18, respectively. (**B**) Overlap between previously identified human genomic DNA G4 regions (*20*) and the set of 6,849 RG4 regions that were identified through RT-stop profiling of RNA from either HeLa or HEK293T cells, uniquely mapped to the reference genome, and did not overlap with each other. (**C**) Contiguous G-runs in RG4 regions. For continuous G-runs of the minimal lengths indicated (G_2 to G_5), the distribution of RG4 regions with the indicated numbers of runs is plotted (red), counting all non-overlapping runs (separated by ≥ 1 nucleotide) within 60 nt upstream of each RT stop. For comparison, the analysis was repeated with a control cohort of 60-nt mRNA regions with equivalent G content (black). For each of the four G-run lengths, RG4 regions had more continuous G-runs than did control regions ($p < 10^{-15}$, one-sided K-S test).

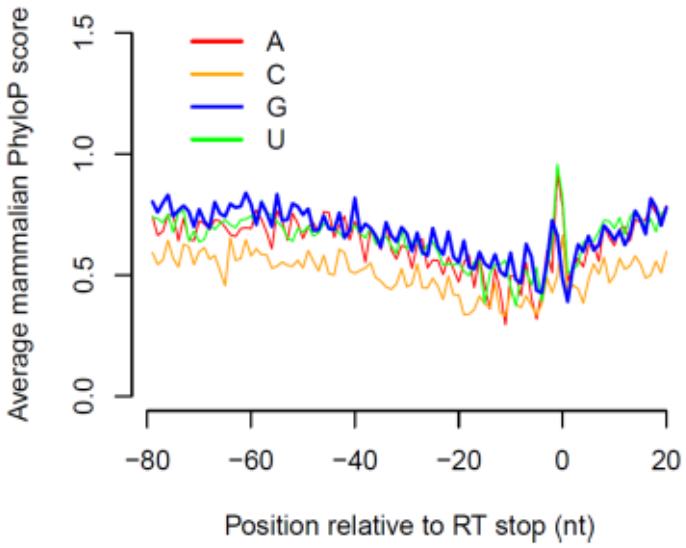

Fig. S5. Sequence conservation of the 6,857 RG4 regions identified in mESCs (table S1). For each of the four nucleotides, PhyloP scores (*40*) from the mouse-centric whole-genome alignments (UCSC Genome Browser, version 2014-07-24) were averaged at each position relative to K^+-dependent strong RT stops. Overall, scores of the G residues resembled those of the A and U residues, and scores within the RG4 regions (positions –60 – 0) resembled those of the flanking regions.

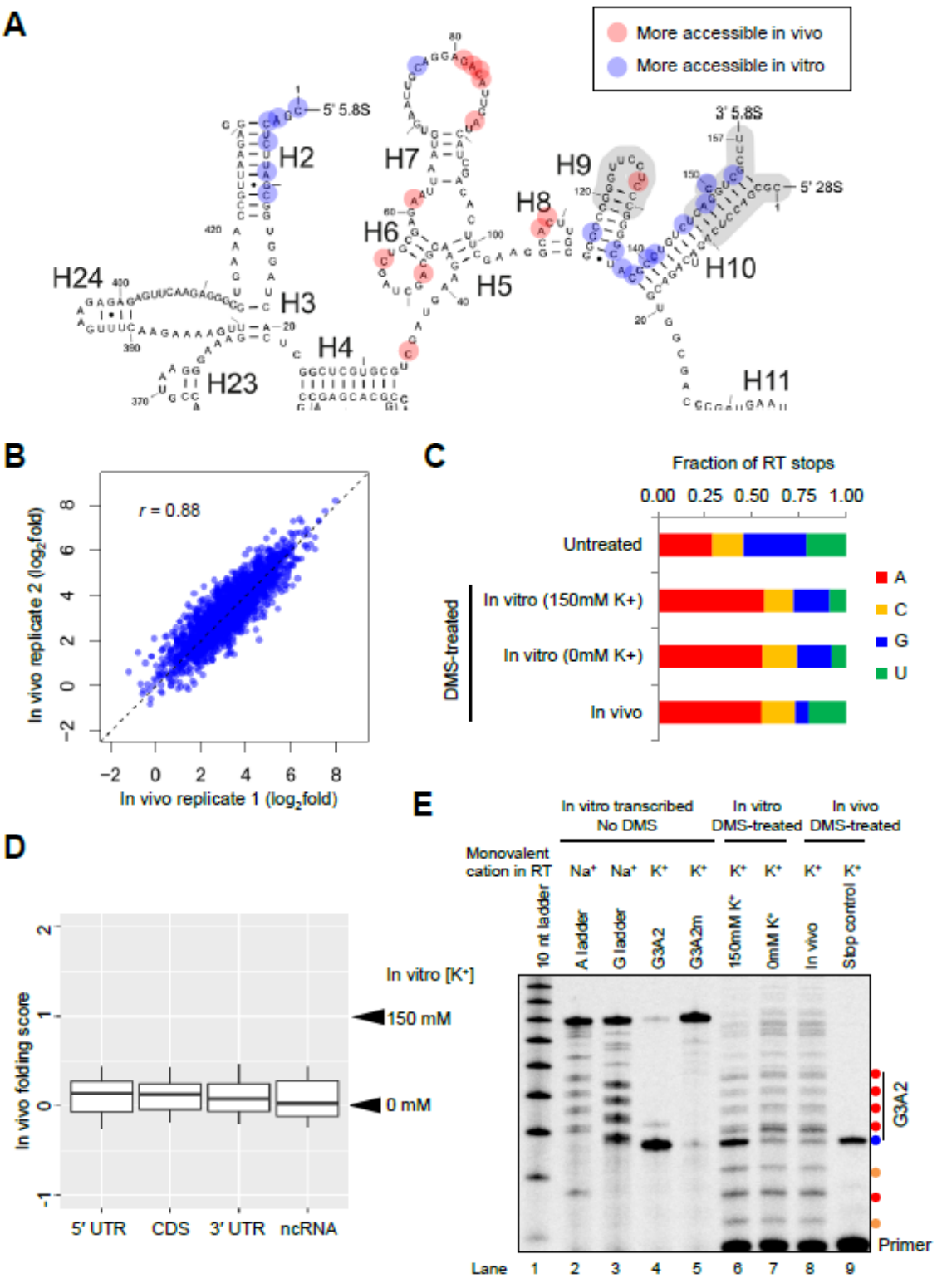

Fig. S6. In vivo DMS probing (*legend on next page*).

Fig. S6. In vivo DMS probing. (**A**) RT stops at A and C residues of 5.8S rRNA, comparing the results of treating ESCs with DMS to those of treating purified refolded RNA. Nucleotides that were ≥2 fold more accessible in vivo or in vitro are labeled in red or blue, respectively. (**B**) A comparison of fold-enrichment values from two biological replicates. For each replicate, mESCs were treated with DMS, and the mRNA was extracted and RT-stop profiled. Pearson's correlation coefficient (*r*) is shown. The data from these to replicates were combined in the following analyses. (**C**) The nucleotide identities at RT-stop reads from the untreated and DMS-treated samples. Note that the fraction of reads that stopped at A and C were similar across all DMS-treated samples. (**D**) In vivo folding of RG4 regions mapping to non-coding RNA (ncRNA) or different regions of mRNAs. Box plots indicate the distributions of in vivo folding scores (median, line; box, quartiles; whiskers 10th and 90th percentiles). In vitro reference levels are indicated. (**E**) The analysis of Fig. 3F, but showing additional lanes with markers and controls. Lanes 6–9 are the ones of Fig. 3F. Also shown is a marker lane (lane 1) and four lanes resolving RT products of in vitro transcribed versions of the RNA that contained the G3A2 quadruplex. Lanes 2 and 3 are A and G ladders, which were generated by including chain terminators (ddTTP or ddCTP, respectively) during RT reactions that were done in the presence of Na^+ instead of K^+. Note that a stop from a chain terminator yields a product that is 1 nt longer than a stop at a RG4 or methylated A or C residue. Lanes 4 and 5 show that the strong RT stop normally observed at the G3A2 quadruplex (lane 4) was abolished when the G3A2 quadruplex was replaced with a mutant version (G3A2m) that contained two G-to-A point substitutions designed to disrupt the quadruplex (lane 5).

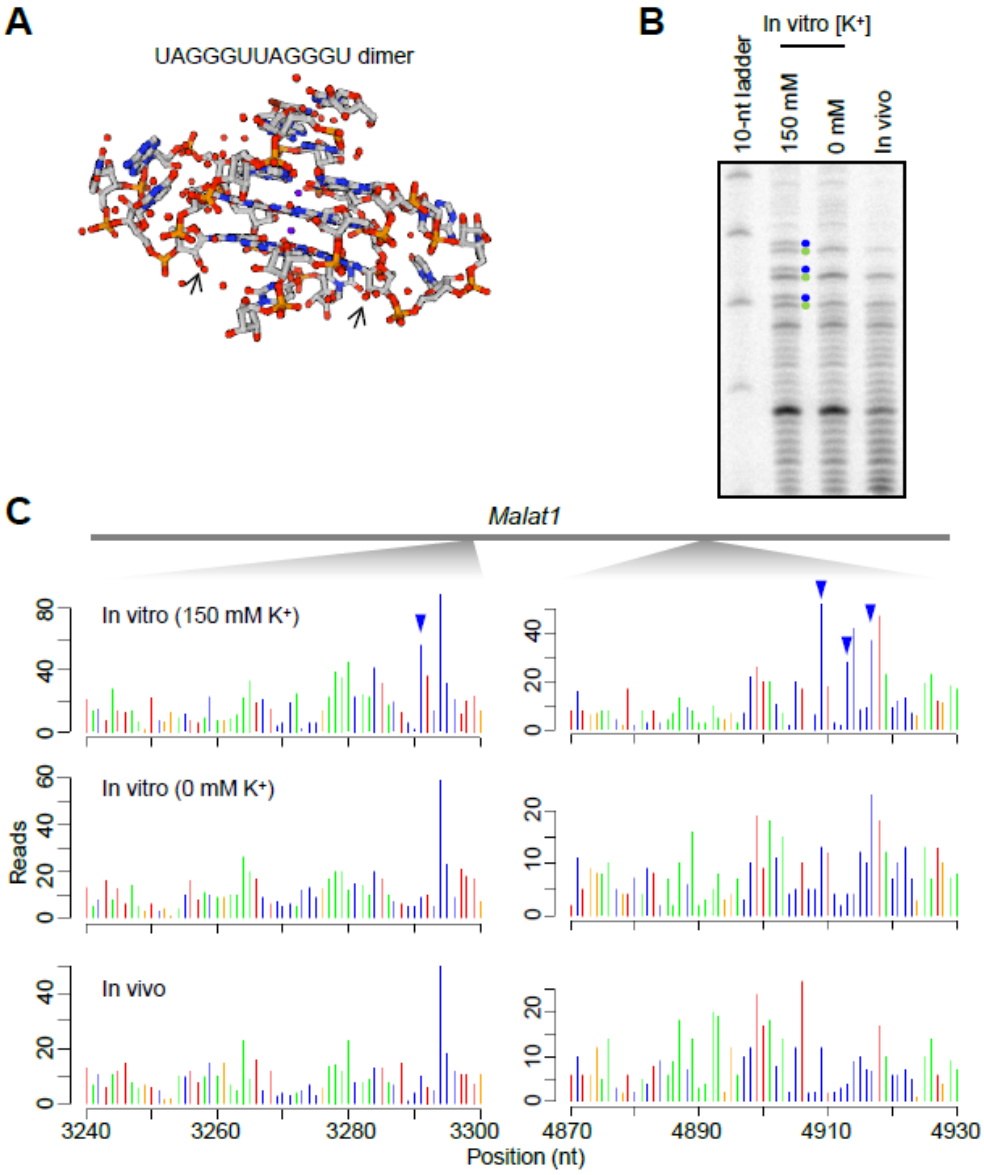

Fig. S7. Exposure of 2′-hydroxyl groups in a parallel RG4 structure. (**A**) The crystal structure of an intermolecular RG4 formed from two molecules of the indicated sequence (PDB accession: 3IBK) (*41*). The exposed 2′-hydroxyl groups of the last G residues of the G tracts that preceded the two loops (UUA segments) are indicated (arrows). (**B**) Gene-specific primer extension of an ectopically expressed G3U quadruplex probed with NAI in vitro (after folding in either 0 or 150 mM K^+) or in HEK293 cells. Shown is a phosphorimage of a denaturing gel that resolved the extension products of a P^{33}-radiolabeled primer. RT stops corresponding to the preferential modifications of NAI within the G3U quadruplex are indicated (blue dots, G residues; green dots, U residues), which included the modifications specific to the folded quadruplex (blue dots). (**C**) RT-stop profiles of two RG4s and their flanking regions within the *Malat1* RNA, showing read counts color-coded according to the identity of the template nucleotide at the stall (position 0). G residues preferentially modified in the presence of 150mM K^+ are indicated (blue arrowheads).

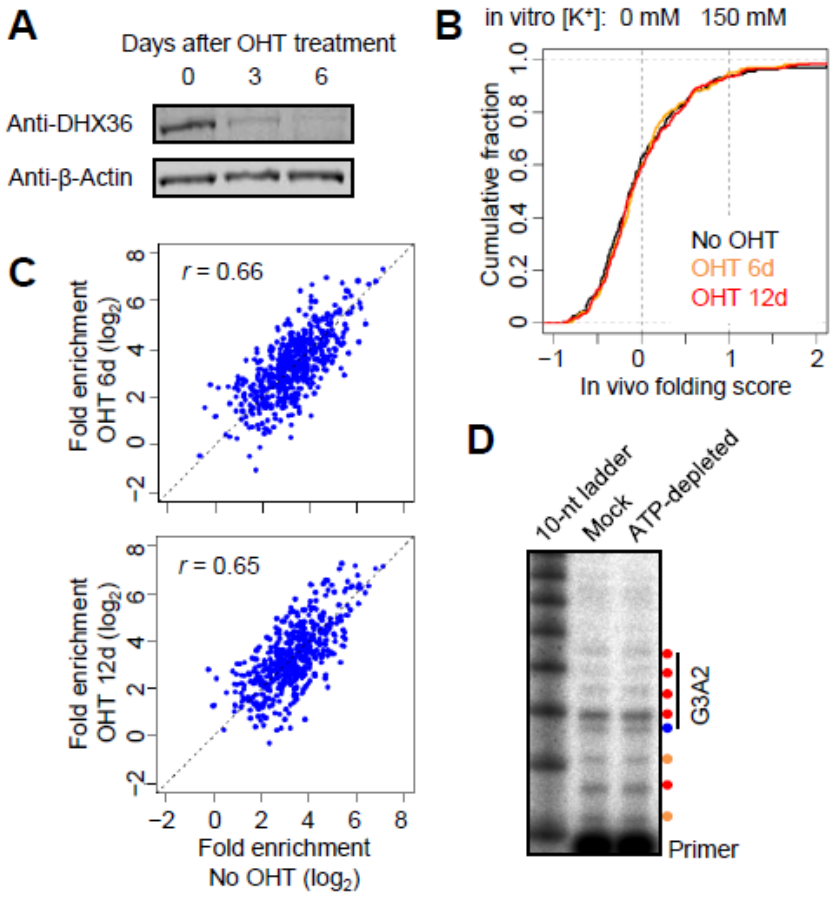

Fig. S8. Robust unfolding of RG4 regions in vivo, despite depletion of either DHX36 or ATP. (**A**) Decrease in DHX36 protein levels in MEFs after adding 1 μM 4-hydroxytamoxifen (OHT), which induces Cre-mediated loss of the *Dhx36* gene. Shown is an immunoblot probed for DHX36 (*top*) and β-Actin (*bottom*), which served as the loading control. After 6 days of OHT treatment, DHX36 protein decreased >95%. (**B**) In vivo folding scores observed after deleting *Dhx36* (6 or 12 days after OHT treatment, OHT 6d and 12d, respectively) compared to those observed after no deletion (no OHT). Shown are distributions of scores for the 268 RG4 regions quantified in all three samples. (**C**) Correspondence between the RT-stop signals observed with and without *Dhx36* deletion. Scatter plots show the fold enrichment values for RT-stop signals of each RG4 region observed after 6 or 12 days of OHT treatment relative to those observed after no treatment (*top* and *bottom*, respectively). Spearman's correlation coefficients are shown (*r*). (**D**) DMS probing of the G3A2 quadruplex in yeast cells either after depleting ATP or after mock depletion. To deplete ATP, exponential-phase yeast were treated with 10 mM sodium azide and 10 mM deoxyglucose for 1 hour. Otherwise, as in Fig. 4D.

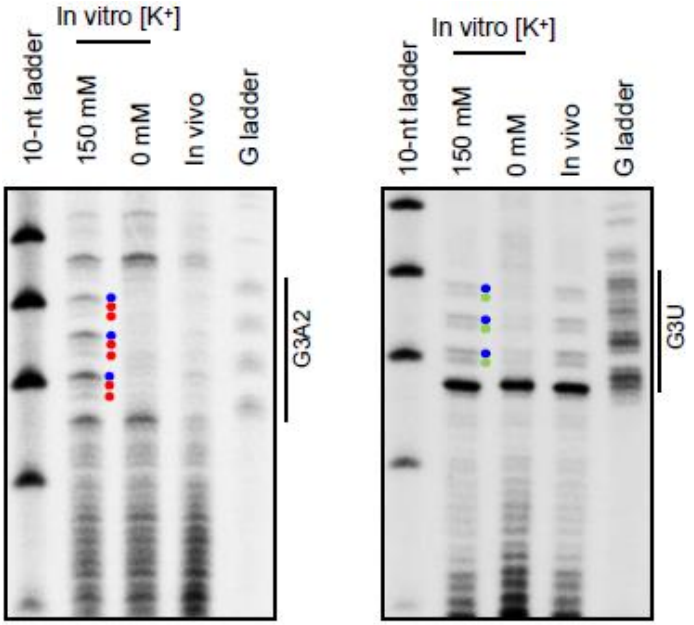

Fig. S9. NAI probing of the G3A2 (*left*) and G3U (*right*) quadruplexes in *E. coli*. Stops at modifications specific to the folded RG4 structures are indicated (blue, G residues; red, A residues; green, U residues). The G ladders were generated as in fig. S6E. The primer was end-labeled with P^{33}. As expected from the results of DMS probing (Fig. 6C), intracellular NAI probing of the G3U quadruplex yielded a strong signal for the pattern of modifications specific to the folded quadruplex. In contrast, intracellular NAI probing of the G3A2 quadruplex yielded only a weak signal for the pattern of modifications specific to the folded quadruplex. This apparent discrepancy with the results of DMS probing for the G3A2 quadruplex (Fig. 6B and C) might be due to either protein binding in vivo that occluded NAI modification at diagnostic 2′-hydroxyl groups, diminishing the signal in this assay, or protein binding in vivo that occluded DMS modification at diagnostic N7 positions of G residues, causing a false positive in the analyses of Fig. 6B and C.

Reference

39. Palenik B, et al., The genome of a motile marine Synechococcus. Nature **424**, 1037-1042 (2003). doi: 10.1038/nature01943; pmid: 12917641
40. K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, A. Siepel, Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res **20**, 110-121 (2010). doi: 10.1101/gr.097857.109; pmid: 19858363
41. G. W. Collie, S. M. Haider, S. Neidle, G. N. Parkinson, Acrystallographic and modelling study of a human telomeric RNA (TERRA) quadruplex. Nucleic Acids Res. **38**, 5569–5580 (2010). doi: 10.1093/nar/gkq259; pmid: 20413582

Additional Data table S1 (separate file)

Strong RT stops of mESCs and identification of RG4 regions.

Additional Data table S2 (separate file)

RG4 regions in the HEK293T and HeLa transcriptomes.

Additional Data table S3 (separate file)

Quantification of in vitro and in vivo folding of RG4 regions in mESCs.

Additional Data table S4 (separate file)

Identification and DMS probing of RG4 regions in *S. cerevisiae*.

Additional Data table S5 (separate file)

Quantification of NAI modifications of RG4 regions in mESCs.

Additional Data table S6 (separate file)

DMS probing of RG4 regions in bacteria.