# Transcriptome-wide profiling of RNA G-quadruplex structures using rG4-seq

**Shankar Balasubramanian** ( sb10031@cam.ac.uk )
Department of Chemistry, University of Cambridge, Cambridge, UK. Cancer Research UK, Cambridge Institute, Cambridge, UK.

**Chun Kit Kwok**
Department of Chemistry, University of Cambridge, Cambridge, UK. Cancer Research UK, Cambridge Institute, Cambridge, UK.

**Giovanni Marsico**
Department of Chemistry, University of Cambridge, Cambridge, UK. Cancer Research UK, Cambridge Institute, Cambridge, UK.

**Aleksandr B. Sahakyan**
Department of Chemistry, University of Cambridge, Cambridge, UK. Cancer Research UK, Cambridge Institute, Cambridge, UK.

**Vicki S. Chambers**
Department of Chemistry, University of Cambridge, Cambridge, UK. Cancer Research UK, Cambridge Institute, Cambridge, UK.

---

**Method Article**

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

**Posted Date:** September 20th, 2016

**DOI:** https://doi.org/10.1038/protex.2016.060



---

# Abstract

Alternative RNA structures such as RNA G-quadruplexes \(rG4s) are often important in gene regulation and cellular processes, however, there is no experimental method for transcriptome-wide assessment of rG4 structures. Here, we introduce a novel approach called rG4-seq, which marries rG4-mediated reverse transcriptase stalling with next-generation sequencing to enable in vitro profiling of rG4 structures on a transcriptomic scale at nucleotide resolution. Upon high-throughput sequencing and bioinformatic analysis, the structural features and distribution of rG4s could be determined, as recently reported using extracted polyadenylated-enriched RNA from HeLa cells \(Kwok et al, 2016). rG4-seq is readily applicable to any transcriptome, allowing global studies of rG4 structures and their potential regulatory roles in biology.

# Procedure

**rG4-seq experimental protocol** _Human cell culture and total RNA preparation_ 1. Culture authenticated human HeLa cells with no mycoplasma contamination in DMEM media \(Sigma, D6429) supplemented with 10% fetal bovine serum \(Sigma, 16140071). 2. Wash the cells \(70-80% confluence level) with Phosphate-buffered saline \(PBS, D8537). 3. Detach the cells from surface using trypsin-EDTA \(Gibco, 25200056). 4. Inactivate Trypsin by DMEM media supplemented with 10% FBS. 5. Pellet the suspended cells and extract total RNA using Qiagen RNeasy Plus Mini Kit \(74134). following manufacturer's protocol. 6. Remove the genomic DNA during the RNA extraction process by gDNA eliminator columns provided in Qiagen RNeasy Plus Mini Kit. _rG4-seq library preparation_ 7. Use 300 µg of total RNA per polyA purist kit reaction \(Ambion, AM1922), which normally yield ~1.5 µg after two rounds of polyA selection following manufacturer's protocol. 8. Check the polyadenylated-enriched RNA with Agilent 2200 tapestation to look for reduction of rRNA peak as compared to total RNA input. 9. Perform RNA random fragmentation 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. 10. Use RNA clean and concentrator \(Zymo research, R1016) according to manufacturer's protocol to clean and concentrate the fragmented RNA. 11. Check the RNA with tapestation to determine the size of the fragmented RNA. 12. Perform 3' dephosphorylation using 8 µl sample, 1 µl 10× T4 PNK buffer, 1 µl T4 PNK enzyme \(NEB, M0201L) with no ATP added at 37°C for 30 min. 13. Conduct 3'adapter ligation 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, M0351L) at 25°C for an hour. 14. Use RNA clean and concentrator according to manufacturer's protocol. 15. Divide the sample into three parts \(10 µl each) for 150 mM $Li^+$, 150 mM $K^+$, and 150 mM $K^+$+ 5 µM PDS conditions for reverse transcription. 16. Add 1 µl of 10 µM unlabelled reverse primer \(5'-CAGACGTGTGCTCTTCCGATCT-3'), and 6 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). 17. Heat the mixture at 95°C for 1.5 min and cool it at 4°C for 1.5 min, then 37°C for 15 min for system equilibration. 18. At the beginning of the 37°C incubation, add 2 µl of nuclease-free water or 50 µM of PDS to the reaction and mix thoroughly. 19. After the 15 min incubation,

add 1 µl of Superscript III \(200U/µL) \(ThermoFisher Scientific, 18080085) and carry out the reverse transcription at 37°C for 50 min. 20. Add 1 µl of 2M NaOH and degrade RNA at 95°C for 10 min. 21. Purify the cDNAs by 10% denaturing TBE gel \(Novex, EC6875BOX) and collect the size ~35-500nt. 22. Crush and soak the gel in 1×TEN250 and incubate it at 80°C for 30 min 23. Use RNA clean and concentrator following manufacturer's protocol. 24. Add 1 µl 50 µM 5'adapter \
(5'/5Phos/AGATCGGAAGAGCGTCGTGTAGCTCTTCCGATCTNNNNNN/3SpC3/3') to the purified cDNAs \
(8 µl). 25. Heat the sample at 95°C for 3min, and then cool it to room temperature 26. Add 10 µl of 2× Quick T4 ligase buffer and 1 µl Quick T4 DNA ligase \(NEB, M2200L) and incubate at 20°C overnight. 27. Purify the ligated cDNAs by 10% denaturing TBE gel and obtain the size 75-500nt, followed by gel extraction step as described in steps 22-23. 28. Perform PCR reaction \(25 µl) 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' AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT 3') and 1 µl 10 µM reverse primer \(e.g. index 2) \(5' CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT 3'), 10.5 µl template and 12.5 µl 2× KAPA HiFi readymix \(Kapa, KK2602). 29. Purify the amplified libraries with 1.8% agarose gel, and obtain the size 150-500 bp. 30. Extract the DNA libraries with Genejet gel extraction kit \
(ThermoFisher Scientific, K0691) following manufacturer's protocol. 31. Perform qPCR on the purified libraries with KAPA Universal Quant Kit \(Kapa, KK4824), and pooled the libraries accordingly. 32. Submit the pooled libraries for next generation sequencing on NextSeq500 machine \(Illumina) for 1 × 150 bp cycle run. **rG4-seq computational analysis** _Sequencing pre-processing_ 33. Trim the fastq files generated by sequencing \(150 bp, single-end) by the trim galore software for removal of Illumina sequencing adapters and low quality tails \
(http://www.bioinformatics.babraham.ac.uk/projects/trim_galore). 34. Align the trimmed data to the human reference genome version hg19 by using the tophat2 software \
(https://ccb.jhu.edu/software/tophat/index.shtml). Download the human genome sequence and gene annotation files for the alignment from the Illumina iGenomes support website \
(https://support.illumina.com/sequencing/sequencing_software/igenome.html). 35. Remove the aligned reads with mapping quality below 30 using the samtools \(http://samtools.sourceforge.net). Calculate the coverage bedGraph files using the bedtools \(http://bedtools.readthedocs.org/), use exon features merged by gene as interval file. _Scoring RT stalling events_ 36. Load coverage files for processing in R \
(https://www.r-project.org/) and convolve the coverage signal with a step-like filter of order 10 to highlight drops in signal at each base \(R 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]). 37. Convolve the coverage signal with a different step-like filter of order 10 to normalize for the total coverage upstream of each base \(R 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]). 38. Calculate the ratio of the two convolved signal 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. 39. Calculate local maxima of the normalized convolved signal, indicating genomic locations where the coverage drop is most pronounced. Assign the signal at those locations a RTS \(reverse transcriptase stalling) value. 40. For each replicate and for each local

maxima \(identified as described above), remove bases with single-base coverage below 6 and coverage drop signal \(RTS) below 0.2 \(20% of reads stalling) to eliminate low-confidence data points and reduce the number of statistical tests performed in the following step. 41. Select two conditions, for instance $K^+$ and $Li^+$, and use the normalized convolved coverage signal for both to fit a linear model \(function lm in R) and estimate the p-value of the fitting through ANOVA testing. 42. correct all p-values from the linear models contrasting any two conditions for multiple hypothesis testing by applying FDR correction on all tested local maxima; 43. Identify significant regions as those having FDR≤0.1 and refer to as scoring regions, or “RTS sites”, yielding a value in the range 0-1, where 0 no stalling and 1 full stalling. Remove sites with RTS value<0.25 and not overlapping any other sites \(cases of overlapping G-quadruplex structures or multiple structural isoforms) from the analysis to further remove regions with subtle effects. Assignment to transcripts. 44. Convert the genomic coordinates of RTS regions to transcript coordinates, according to the following steps: • Calculate FPKM \(Fragments Per Kilobase per Million of mapped reads) for each transcript isoform in the $Li^+$ condition using the software cufflinks \(http://cole-trapnell-lab.github.io/cufflinks/) and consider transcripts with FPKM≥0.5 as expressed, resulting in 17,622 transcript isoforms belonging to 12,300 unique genes; • map genomic scoring regions 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, assess the second most expressed transcript and so on; • calculate transcript coordinates and sequence for the mapped scoring regions to account for intron skipping. _Hierarchical assignment of rG4s_ 45. Extract sequences extending from the RTS stalling site to 50 base pairs upstream in the transcript for each RTS site and assign to different rG4s structural sub-classes, defined as follows \(regular expressions used for pattern matching shown in brackets): • G3L1-7, canonical rG4s with loop length between 1-7nt \('\(G3+N1-7)\{3,}G3+', 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., "G3+N8-12G3+N1-7G3+N1-7G3+" or 'G3+N1-7G3+N13-21G3+N1-7G3+'); • Bulges, rG4s with a bulge of 1–7nt in one G-tract or multiple 1nt bulges \(e.g., 'G3+N1-9G3+N1-9\(GGH1-7G|GH1-7GG)N1-9G3+' or '\(GGHG|GHGG)N1-9 \(GGHG|GHGG)N1-9G3+N1-9G3+', with H = A, U or C); • 2-quartet, rG4s with 4-tracts of two consecutive Gs \('\(G2+N1-9)\{3,}G2+'); 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. 46. When matching multiple categories, assign a region to the class with higher predicted stability, i.e. \(from first to last), canonical rG4s, long loops, bulges, 2-quartet. _Nucleotide content analysis and delta free energy analysis_ 47. Extract exonic canonical rG4s \(category G3L1-7) with coverage ≥ 20 in the $Li^+$ condition; equally extend each rG4 sequence upstream and downstream to enclose a total region of 90 nt including both the rG4 motif and flanking regions. 48. Count the occurrence of C, CC, CCC, and CCCC motif within each extended region containing detected or undetected rG4s and divide by the region length, yielding a motif density value per sequence. 49. Similarly, count the occurrence of U, A and G motif. 50. Compute the average value of detected and undetected G3L1-7 density and calculate the ratio of the average detected to average undetected density. 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 G3L1-7. _Delta free energy analysis_ 51. For the same G3L1-7 identified in step 47., equally extend G3L1-7 upstream and downstream to enclose a total regions of 90nt including both the rG4 motif

and flanking regions. 52. Calculate the free energy of ensemble of these sequences using the RNAFold software of the ViennaRNA package \(http://www.tbi.univie.ac.at/RNA/) and compare the detected and undetected G3L1-7. _mRNA region \(UTRs and CDS) analysis_ 53. Calculate the overlap of each scoring region \(RTS site) with 5'UTR, CDS \(coding sequences) and 3'UTR \(bedtools intersect). 54. Assign rG4s partially overlapping multiple regions \(e.g., 5'UTR and CDS) to both regions and count twice. 55. Divide the number of overlapping regions to each of the three annotated features by the total region size in base pairs and multiply by 1000, therefore yielding the rG4 density per kilobase \(density per kb). _Average mRNA profile_ 56. Normalize each transcript profile to the same length and divide its 5'UTR, CDS and 3'UTR 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. 57. For each RTS site, assess the belonging to a given bin and compute 1/bin_size \(count normalized by bin size in bp). 58. Average all normalized counts per bin for scored regions in all transcripts and plot. _Regulatory sites analysis_ 59. Obtain microRNA target sites and polyadenylation signal \(PAS) sites from the TargetScan database of predicted miRNA target sites \(http://targetscan.org) and from the GENCODE project \(http://www.gencodegenes.org, release 19) respectively. 60. Transform genomic coordinates of all sites into transcript coordinates in order to calculate distance by skipping introns. 61. Independently for each regulatory feature class and for each scoring region \(RTS site), calculate the distance between the region and the closest feature, and assign 0 for overlapping features. 62. Randomly shuffle the scoring regions by uniform resampling across all the expressed transcripts after merging overlapping exons, in order to avoid over-representation of genes with several alternative isoforms. 63. Build the cumulative distributions of pairwise feature-region distances for RTS sites and random regions. 64. Similarly, assess the cumulative distribution of distance separately for rG4s up- and downstream of the respective regulatory sites. 65. Calculate the fraction of rG4s in proximity \(i.e. $\leq$100 nt) of regulatory sites for all rG4s and for up- and downstream rG4s, and compare to random by using the Chi-squared test for proportions \(function prop.test in R). _RNA structure prediction and PPV comparison_ 66. As some scoring regions \(RTS sites) can be assigned to multiple overlapping genes, extract unique sequences only. 67. Consider overlapping rG4s identified as single rG4s if the G-quadruplex motif assigned through the hierarchical motif analysis \(step 45) coincides, to avoid redundant RTS that would lead to the same structural prediction. 68. Extend the 50 nt scoring regions by 100nt up- and downstream, resulting in regions of 250 nt. 69. Use the RNAFold software \(http://www.tbi.univie.ac.at/RNA/) for the structural prediction of the 250 nt sequences, and repeat the prediction with and without imposing single-strand constraint over the rG4 identified motif. 70. When more than a motif is identified from the structural characterization analysis, impose the single-strand constraint over the G-quadruplex motif closer to the identified stalling site. 71. Use the RNAstructure software \(http://rna.urmc.rochester.edu/RNAstructure.html) to convert dot bracket \(db) files to connectivity table \(ct) files. 72. Perform structure comparison between constrained and unconstrained structural prediction of each scoring region using the function CircleCompare of the same software. 73. Use the PPV \(positive predictive value) returned by CircleCompare to compare strucutres with and without the rG4 constraint. 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_ 74. Consider all the 68 non-human eukaryotic species with genomes and annotations deposited in the Ensembl genome database, as accessed through BioMart \(http://www.ensembl.org). 75. For each rG4, survey the presence of genes in other species that are orthologous to the human genes bearing the corresponding rG4s: for a particular rG4h human \(h) sequence and a non-human \(non-h) species, if an ortholog is present for the rG4h-bearing gene, perform a global sequence alignment of the corresponding human and non-human cDNA sequences. 76. Perform the sequence alignment using the utilities provided in the Biostrings library for R, 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). 77. After the alignment, identify the rG4h sequence in the aligned human cDNAh, using the rG4-seq revealed coordinate of the 3'-end of rG4h, 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). Correct those coordinates to account for the possible gaps in cDNAh sequence introduced in the cDNAh vs. cDNAnon-h alignment. 78. Examine the matched segment from the aligned non-human cDNAnon-h sequence, corresponding to the region in human cDNAh that engulfs the particular rG4h sequence: remove the gaps from that segment \(if introduced during the alignment) and, if a nucleic acid sequence was left, assess the presence of a potential G-quadruplex using the relaxed \(G2+N1-12)3G2+ sequence pattern. 79. Identify the G4 sequences that match with the actual experimentally \(rG4-seq) validated rG4h by their positioning inside the orthologs. Store the outcome of this analysis in a matrix \(N rG4s × 69 species), with 1 if there is a rG4, 0 if there is an ortholog found but without a G4 sequence at the locus matching the human rG4, and NA \(non-assigned) for the cases where even an ortholog is absent. 80. Use the matrix to cluster the data entries in both rG4 and species dimensions, producing a heatmap where the G4 presence patterns are clustered.

# Timing

The rG4-seq experimental protocol takes about 3-4 working days. The rG4-seq computational analysis takes about 2-3 working days.

# Anticipated Results

Application of rG4-seq to the extracted polyadenylated RNA from human HeLa cells was recently reported \(Kwok et al, 2016)[1]. Using this rG4-seq protocol, four biological replicates of each condition \($Li^+$, $K^+$, $K^+$+PDS) were constructed to cDNA libraries. The cDNA libraries were then sequenced on Illumina Nextseq with single-read, 150 cycles. Analysis of the sequencing data reveals widespread formation of rG4 in the human transcriptome.

# References

1. Kwok, CK., Marsico, G., Sahakyan, A., Chambers V.S., Balasubramanian, S. rG4-seq reveals widespread formation of G-quadruplex structures in the human transcriptome. Nat. Methods. \(2016). doi:10.1038/nmeth.3965

## Acknowledgements

This work 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 \(grant no. 099232/z/12/z). C.K.K and G.M contributed equally to this work. We thank members of the Balasubramanian laboratory for comments.