# Supplementary Data

# Preferential formation of Z-RNA over intercalated motifs in long non-coding RNA

Uditi Bhatt[1], Anne Cucchiarini[2], Yu Luo[2], Cameron W. Evans[1], Jean-Louis Mergny[2], K. Swaminathan Iyer[1], and Nicole M. Smith[1,*]

[1] School of Molecular Sciences, The University of Western Australia, Crawley, WA, 6009, Australia.

[2] Laboratoire d'Optique et Biosciences, École Polytechnique, CNRS, INSERM, Institut Polytechnique de Paris, 91128 Palaiseau, France.

* To whom correspondence should be addressed. Tel: +61 8 6488 4423; Email: nicole.smith@uwa.edu.au

## Table of Contents

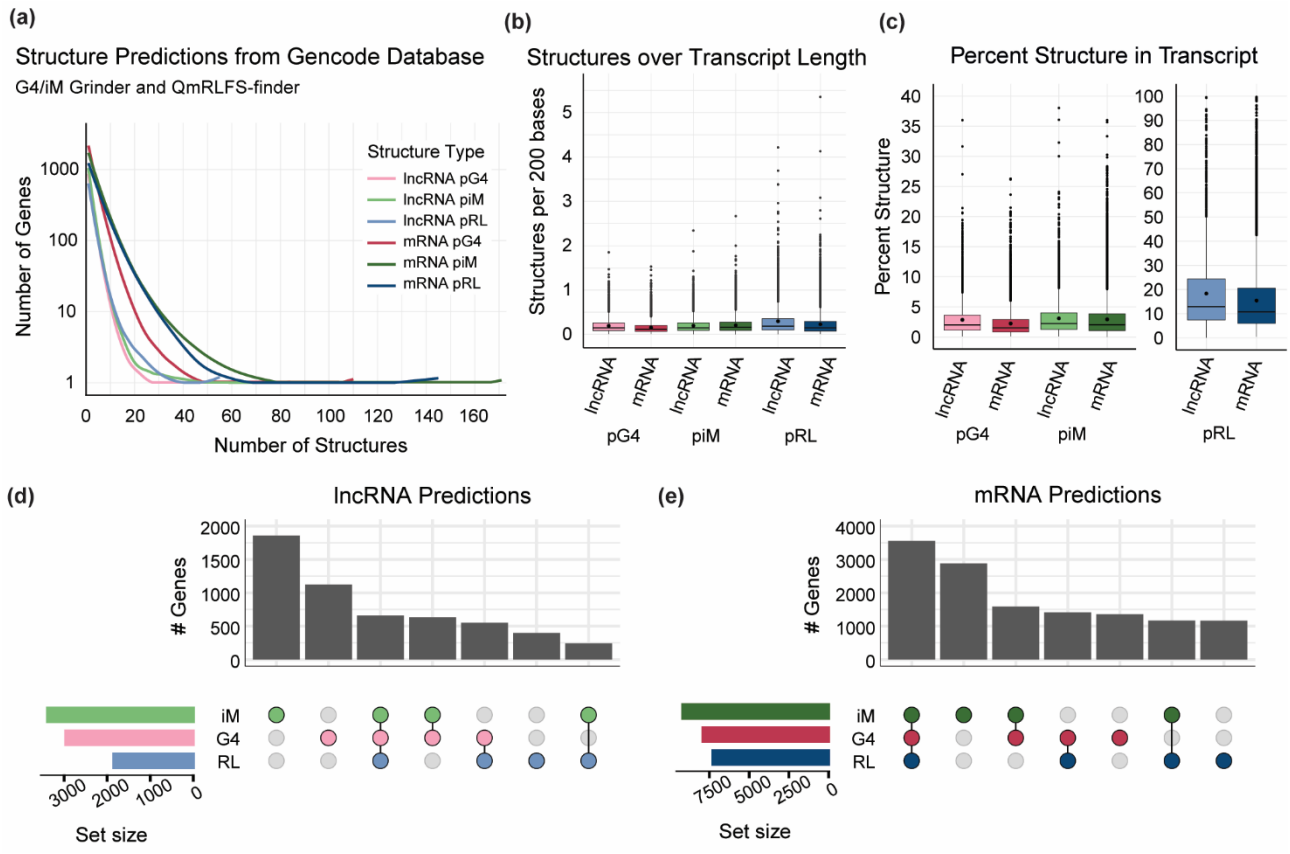



**Figure S1. Unfiltered data of predicted non-canonical secondary structures in lncRNA and mRNA. (a)** Density of each predicted structure type in lncRNA and mRNA genes. **(b)** Box plot showing the distribution of the number of structures over transcript length for each type of predicted structure in lncRNA and mRNA with outliers. **(c)** Box plot showing the distribution of the total percentage of nucleotides in the transcript involved in structure formation for each structure type in lncRNA and mRNA including outliers. **(d)** lncRNA and **(e)** mRNA structure prediction datasets. Total number of each structure type shown in horizontal bar graph 'Set size'.

**Table S1. Total number and percentage of pG4, piM, and pRL structures from lncRNA and mRNA with near-equivalent transcript lengths and GC content.** 14,637 transcripts from 5,254 genes were included for lncRNA and 13,695 transcripts from 6,260 genes were included for mRNA.

| | Structure type | Total predicted structures | Unique genes | % of input genes | Unique transcripts | % of input transcripts | Unique motifs |
|---|---|---|---|---|---|---|---|
| lncRNA | G4 | 805 | 528 | 10.05 | 767 | 5.24 | 584 |
| | iM | 1221 | 646 | 12.30 | 1151 | 7.86 | 703 |
| | RL | 240 | 160 | 3.05 | 192 | 1.31 | 212 |
| mRNA | G4 | 825 | 537 | 8.58 | 776 | 5.67 | 601 |
| | iM | 1003 | 589 | 9.41 | 907 | 6.62 | 665 |
| | RL | 946 | 433 | 6.92 | 613 | 4.48 | 731 |

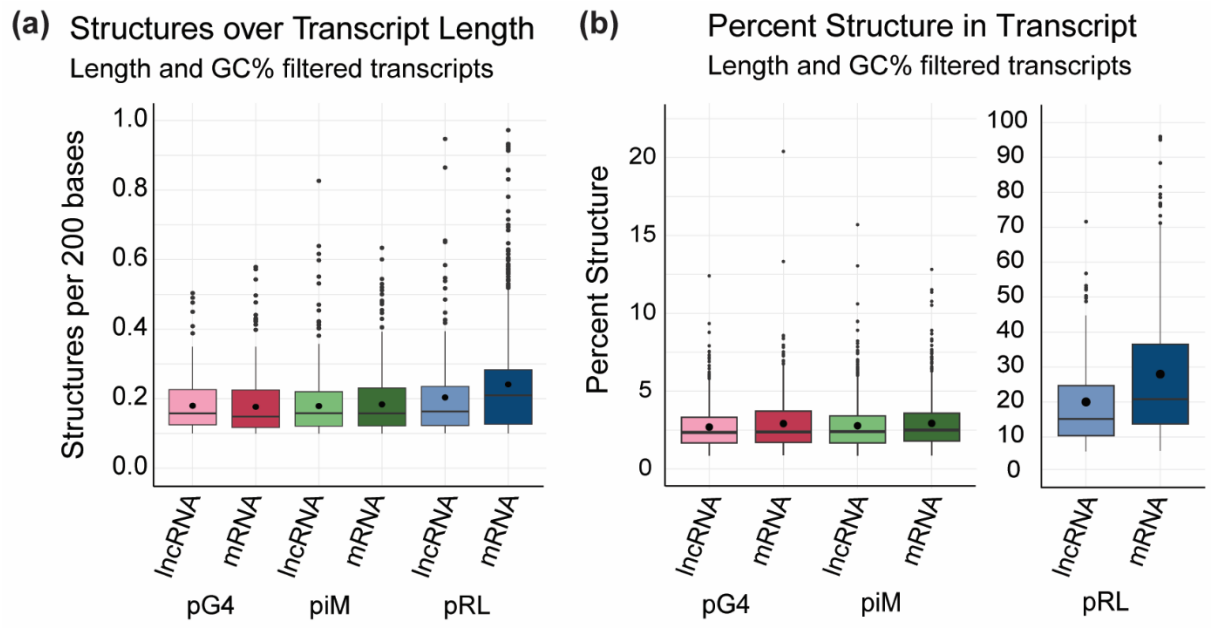
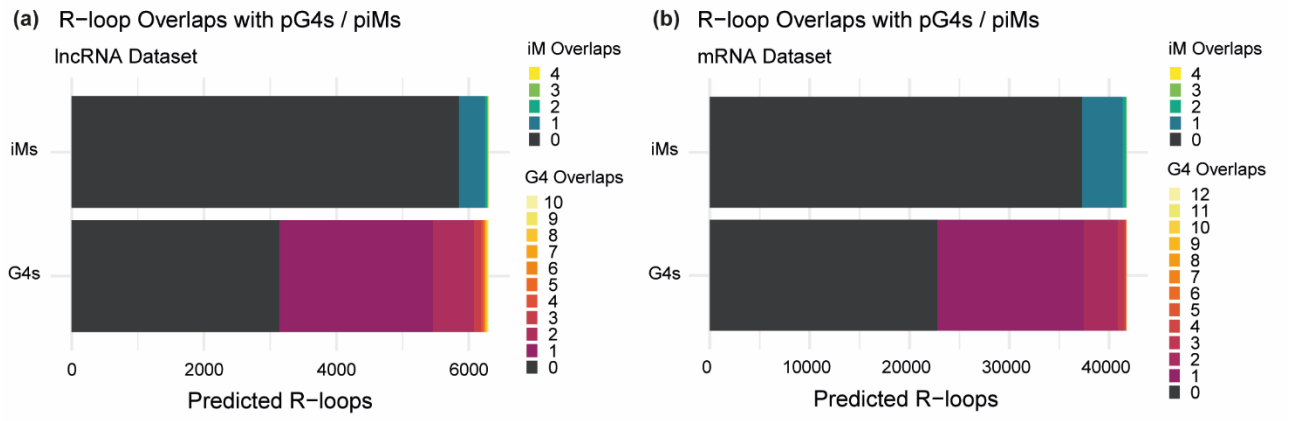



**Figure S2. Unfiltered structure density and percent of transcript data for similar length and GC% lncRNA and mRNA dataset. (a)** Box plot showing the distribution of the number of structures over transcript length for each type of predicted structure in lncRNA and mRNA with outliers. **(b)** Box plot showing the distribution of the total percentage of nucleotides in the transcript involved in structure formation for each structure type in lncRNA and mRNA including outliers.




(c) pRL-pG4 Overlaps

| # Overlapping Nucleotides | # lncRNAs | Percent |
|---|---|---|
| 0 | 3141 | 49.87 |
| 1 | 2315 | 36.75 |
| 2 | 623 | 9.89 |
| 3 | 107 | 1.70 |
| 4 | 30 | 0.48 |
| 5 | 15 | 0.24 |
| 6 | 18 | 0.29 |
| 7 | 6 | 0.10 |
| 8 | 22 | 0.35 |
| 9 | 10 | 0.16 |
| 10 | 12 | 0.19 |

pRL-piM Overlaps

| # Overlapping Nucleotides | # lncRNAs | Percent |
|---|---|---|
| 0 | 5849 | 92.86 |
| 1 | 398 | 6.32 |
| 2 | 32 | 0.51 |
| 3 | 16 | 0.25 |
| 4 | 4 | 0.06 |

(d) pRL-pG4 Overlaps

| # Overlapping Nucleotides | # mRNAs | Percent |
|---|---|---|
| 0 | 22806 | 54.60 |
| 1 | 14633 | 35.03 |
| 2 | 3421 | 8.19 |
| 3 | 642 | 1.54 |
| 4 | 168 | 0.40 |
| 5 | 41 | 0.10 |
| 6 | 21 | 0.05 |
| 7 | 11 | 0.03 |
| 8 | 1 | 0.00 |
| 9 | 6 | 0.01 |
| 10 | 5 | 0.01 |
| 11 | 10 | 0.02 |
| 12 | 5 | 0.01 |

pRL-piM Overlaps

| # Overlapping Nucleotides | # lncRNAs | Percent |
|---|---|---|
| 0 | 37295 | 89.29 |
| 1 | 4083 | 9.77 |
| 2 | 340 | 0.81 |
| 3 | 42 | 0.10 |
| 4 | 10 | 0.02 |

**Figure S3. Sequence overlaps between predicted R-loops and predicted G4s and iMs.** Stacked bar graphs showing pRL sequence overlaps with piM and pG4 sequences in **(a)** lncRNA and **(b)** mRNA. Colour legends represent the number of overlapping nucleotides from grey (lowest) to yellow (highest). Table with number and percentage of RNAs with given number of overlapping nucleotides between pRL and pG4/piM structures for **(c)** lncRNAs and (**d)** mRNAs.

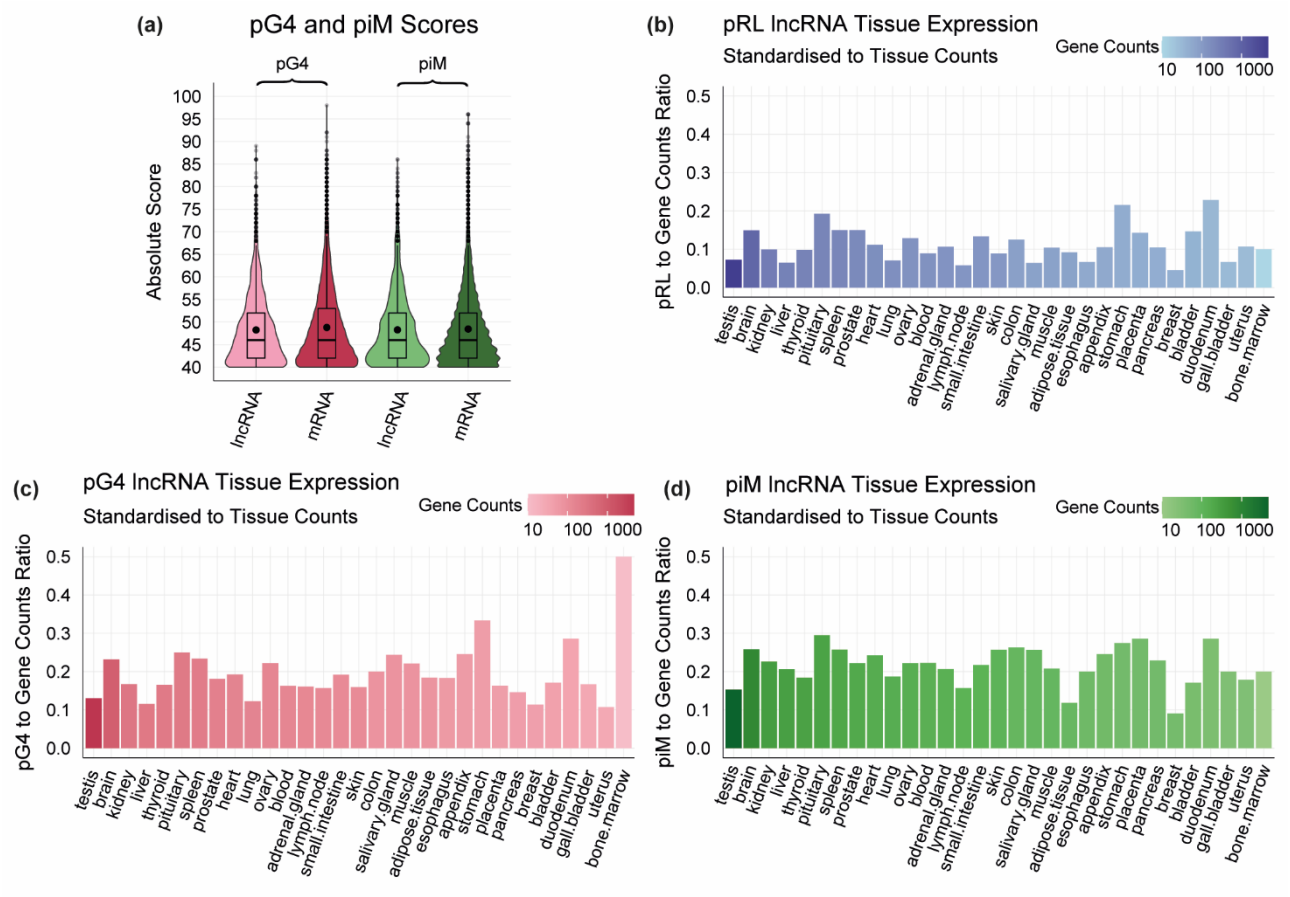



**Figure S4. Score and tissue expression data for pG4s and piMs in lncRNA.** **(a)** Violin + boxplot showing distribution of |score| values for pG4 and piM structures in lncRNA and mRNA. Ratio of number of lncRNA genes with **(b)** pRL, **(c)** pG4, or **(d)** piM structures in a given tissue standardized to the number of genes in the tissue from the initial tissue dataset. Colour gradient indicates the number of genes in the tissue, light = low, dark = high. Tissues with less than 10 genes in the initial dataset were removed.

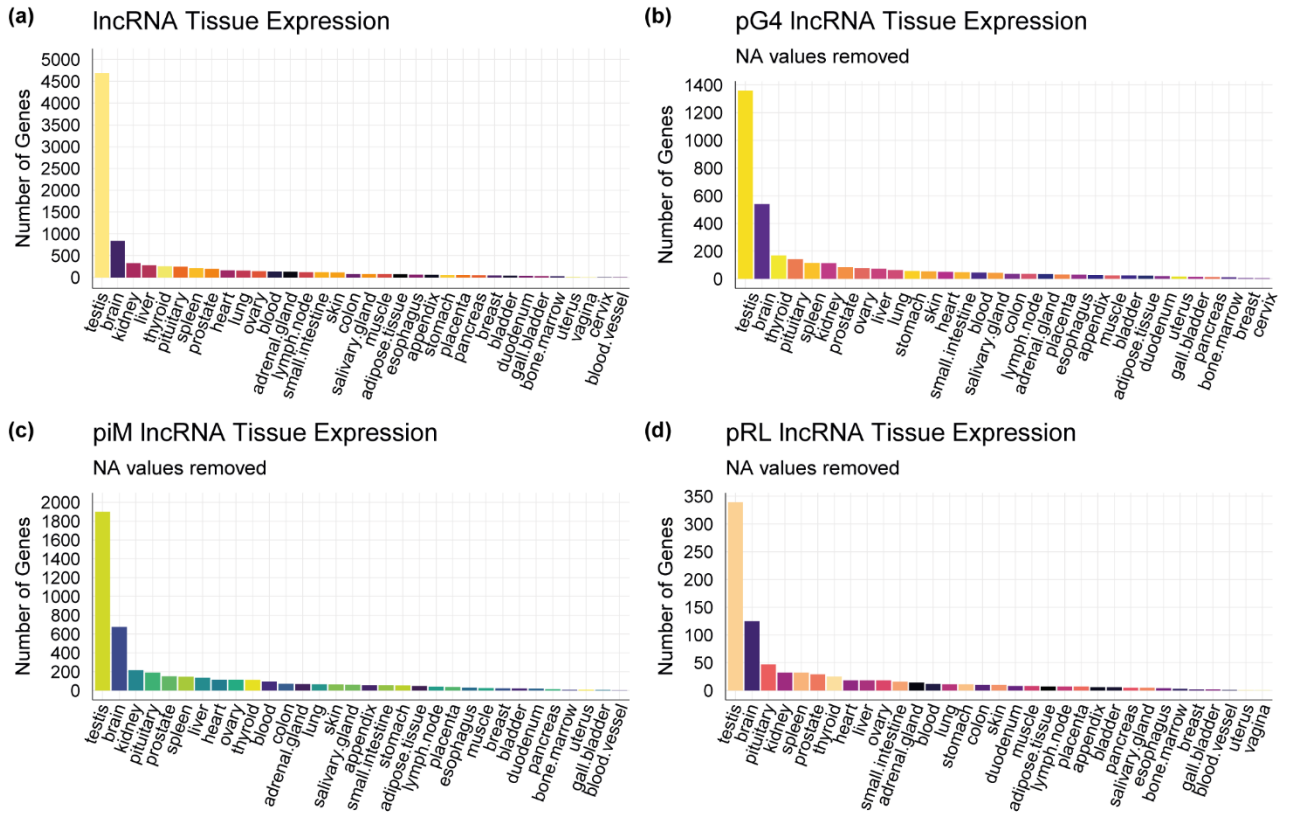



**Figure S5. lncRNA tissue expression.** **(a)** Total number of lncRNA genes per tissue in the input dataset without predicted structures. Total number of lncRNAs genes per tissue with predicted **(b)** G4, **(c)** iM, **(d)** and RL structures.

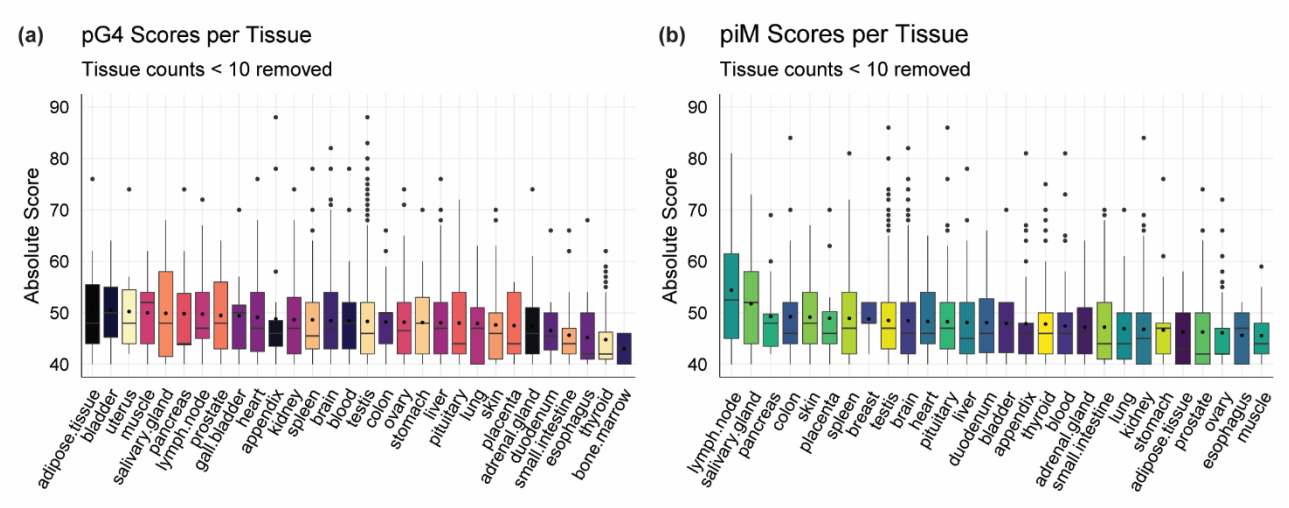
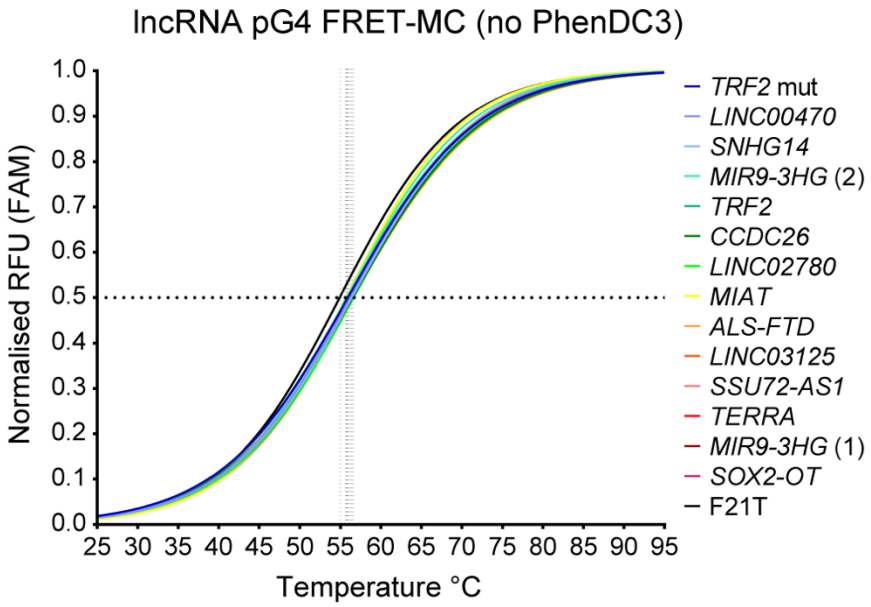



**Figure S6. Score distribution in different tissues.** Boxplots showing score distributions in each lncRNA tissue for **(a)** pG4 structures and **(b)** piM structures. Tissues with less than 10 genes in the initial dataset were removed.

(a)




(b)

| Oligo | *Tm* of F21T (no PhenDC3) |
|---|---|
| F21T only | 55.0 |
| *MIAT* | 55.7 |
| *MIR9-3HG* (2) | 55.8 |
| *TRF2* | 55.9 |
| *TRF2* mut | 56.0 |
| *LINC00470* | 56.2 |
| *SNHG14* | 56.4 |
| *MIR9-3HG* (1) | 56.4 |
| *LINC03125* | 56.4 |
| *CCDC26* | 56.6 |
| *ALS-FTD* | 56.6 |
| *TERRA* | 56.6 |
| *SSU72-AS1* | 56.6 |
| *SOX2-OT* | 56.7 |
| *LINC02780* | 56.7 |

**Figure S7. FRET-MC control data. (a)** Non-linear fit of normalized F21T melt curves ± test RNA pG4 competitor oligos without PhenDC3. Horizontal dotted line at y = 0.5, vertical dotted lines at $T_m$ of F21T in each sample. **(b)** Table of $T_m$ of F21T in all samples from FRET-MC without PhenDC3.

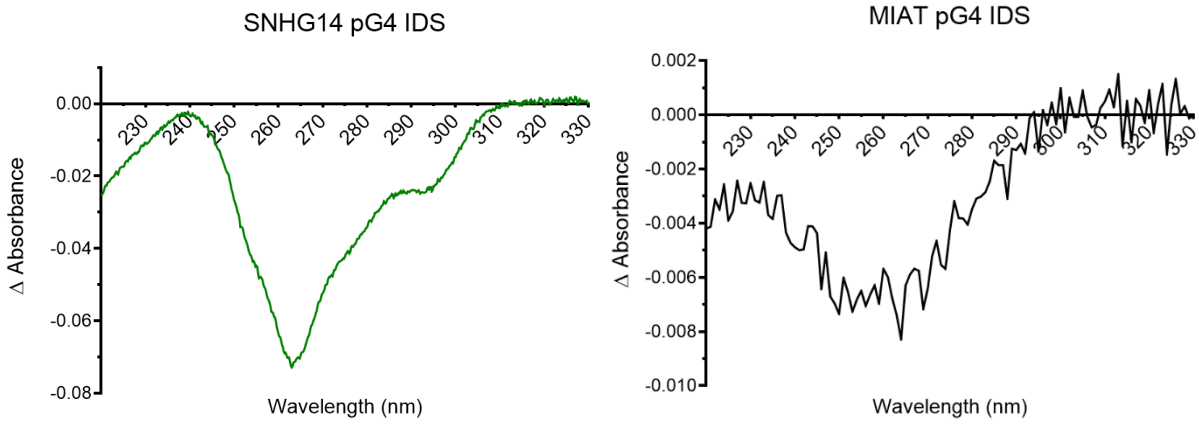
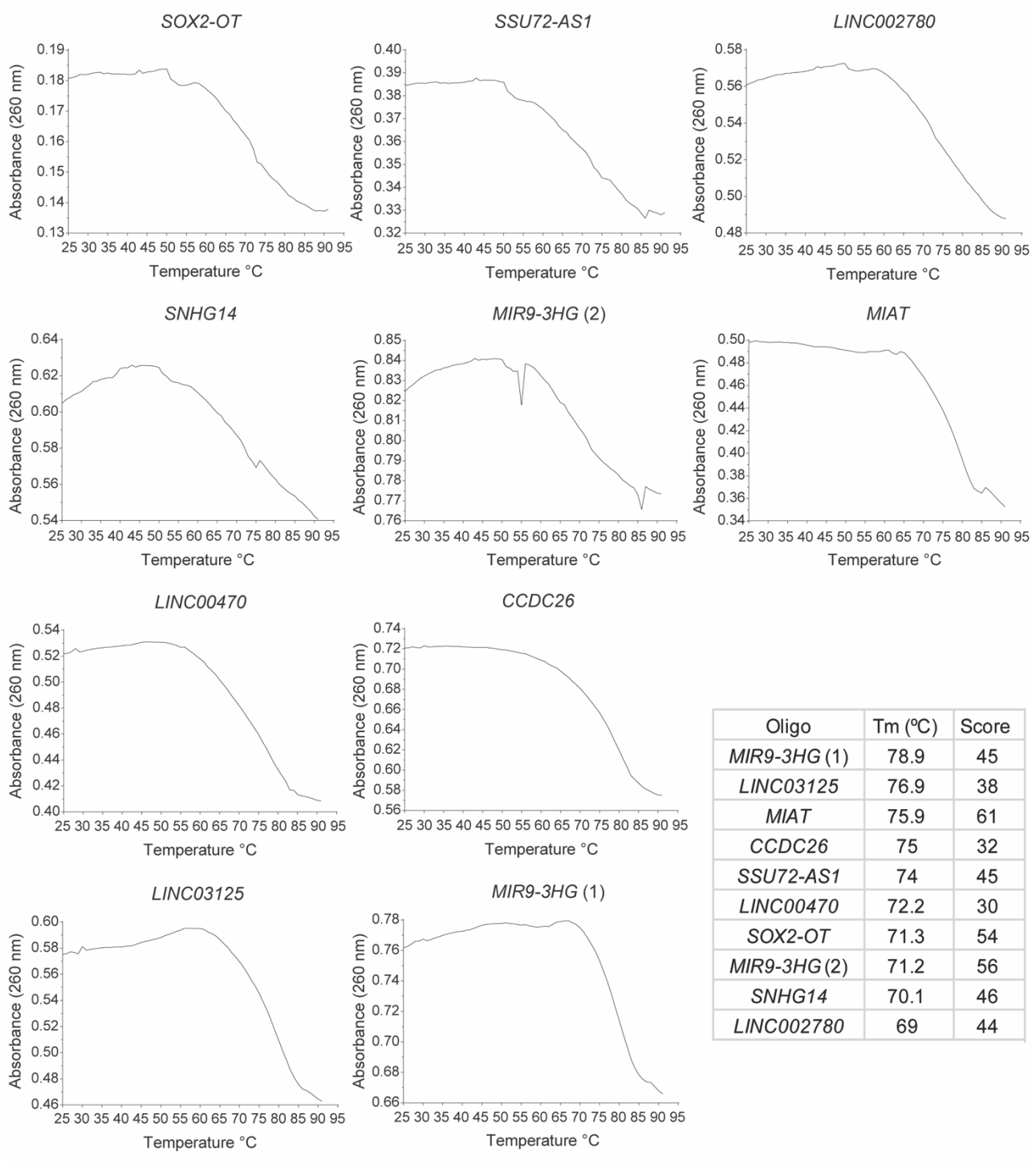



**Figure S8.** Raw IDS data for SNHG14 and MIAT pG4 lncRNA sequences at 37 °C.




| Oligo | Tm (ºC) | Score |
|---|---|---|
| *MIR9-3HG* (1) | 78.9 | 45 |
| *LINC03125* | 76.9 | 38 |
| *MIAT* | 75.9 | 61 |
| *CCDC26* | 75 | 32 |
| *SSU72-AS1* | 74 | 45 |
| *LINC00470* | 72.2 | 30 |
| *SOX2-OT* | 71.3 | 54 |
| *MIR9-3HG* (2) | 71.2 | 56 |
| *SNHG14* | 70.1 | 46 |
| *LINC002780* | 69 | 44 |

**Figure S9.** UV melt curve raw data and table showing $T_m$ for each pG4 lncRNA oligo tested.

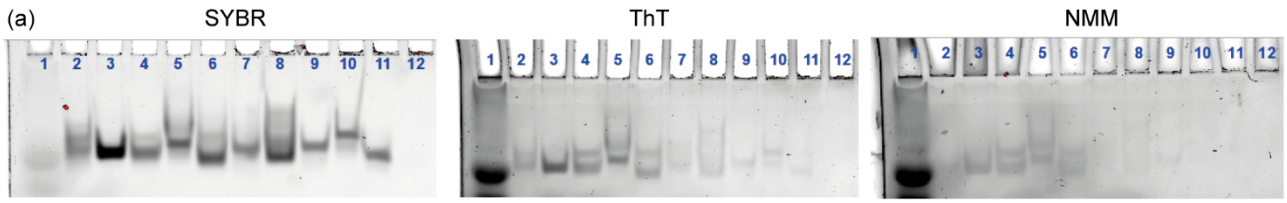



**Figure S10. Polyacrylamide gel for pG4 lncRNA.** 20% Native PAGE for lncRNA pG4 samples stained with **(a)** SYBR Safe, **(b)** ThT, or **(c)** NMM. Lanes – 1: Loading dye, 2: *SNHG14*, 3: *LNC02780*, 4: *SSU72-AS1* 5: *SOX2-OT*, 6: *MIR9-3HG* (2), 7: *MIAT*, 8: *LINC00470*, 9: *CCDC26*, 10: *LINC03125*, 11: *MIR9-3HG* (1).

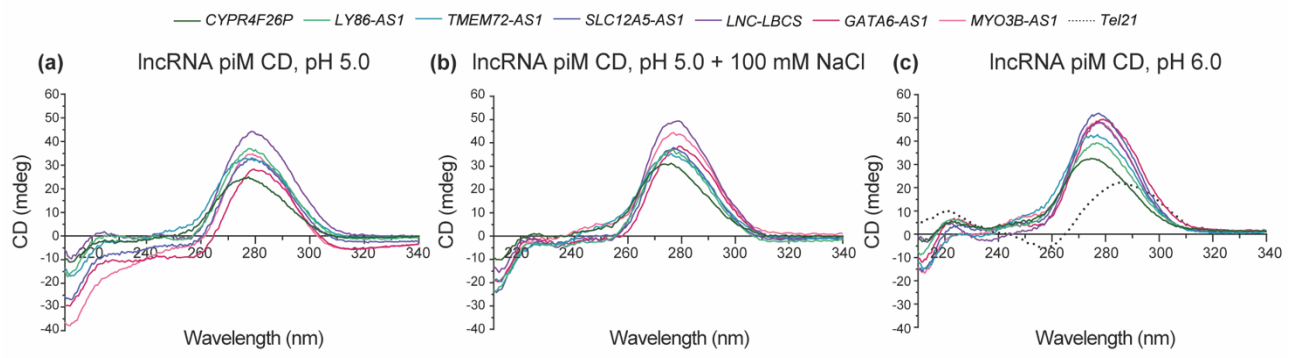



**Figure S11. CD of lncRNA piM (Z-RNA) sequences. (a)** 10 mM NaCaco, pH 5.0, **(b)** 10 mM NaCaco + 100 mM NaCl, pH 5.0, **(c)** 10 mM NaCaco, pH 6.0.

**Table S2.** pG4 sequences selected for biophysical analysis that are present in multiple lncRNAs. Name by which the sequence is referred to in the main manuscript is in bold.

| 5’ -**GGGAGG**CU**GAGGUGGG**-3’ | | | | | |
|---|---|---|---|---|---|
| ***LINC02780*** | *AC021148.3* | *AC130895.1* | *AZIN1-AS1* | *LINC00937* | *SHANK2-AS3* |
| *A1BG-AS1* | *AC022558.3* | *AC135012.1* | *BMP8B-AS1* | *LINC01476* | *SNHG5* |
| *AC003681.1* | *AC022596.2* | *AC138904.3* | *BSN-DT* | *LINC01511* | *SPACA6P-AS* |
| *AC004039.1* | *AC023347.1* | *AL022323.2* | *C12orf77* | *LINC01695* | *THAP9-AS1* |
| *AC004494.1* | *AC024588.1* | *AL022341.2* | *C21orf62-AS1* | *LINC02018* | *TMED2-DT* |
| *AC005005.4* | *AC026362.2* | *AL023882.1* | *CCDC15-DT* | *LINC02021* | *TMEM161B-AS1* |
| *AC006115.2* | *AC027644.3* | *AL109615.4* | *CTBP1-AS* | *LINC02595* | *TMEM254-AS1* |
| *AC006504.5* | *AC041005.1* | *AL157931.1* | *DENND6A-DT* | *LINC02642* | *TMEM51-AS1* |
| *AC007614.1* | *AC067863.1* | *AL359220.1* | *DLGAP1-AS1* | *MCPH1-AS1* | *TSIX* |

| | | | | | |
|---|---|---|---|---|---|
| *AC007922.4* | *AC067930.2* | *AL390071.1* | *EFCAB6-AS1* | *MIR3976HG* | *WAC-AS1* |
| *AC008543.5* | *AC073648.7* | *AL606753.2* | *ENTPD1-AS1* | *MKLN1-AS* | *XACT* |
| *AC008738.4* | *AC073957.3* | *AL627171.4* | *FAM153CP* | *MMP25-AS1* | *XIST* |
| *AC009318.1* | *AC097376.3* | *AL731559.1* | *HCG11* | *NCBP2-AS1* | *Z84486.1* |
| *AC010618.3* | *AC104024.2* | *AL732509.1* | *HELLPAR* | *PART1* | *Z95114.1* |
| *AC010733.1* | *AC107398.3* | *AP001020.1* | *LINC00461* | *PINK1-AS* | *Z98885.3* |
| *AC010907.2* | *AC109635.6* | *AP002518.1* | *LINC00467* | *PRR34* | |
| *AC016687.2* | *AC116158.2* | *AP002847.1* | *LINC00685* | *PSMA3-AS1* | |
| *AC021078.1* | *AC116914.2* | *AP003721.1* | *LINC00836* | *RFPL1S* | |
| 5’-**GGG**C**G**U**GG**U**GG**C**GGG**-3’ | | | | | |
| ***SSU72-AS1*** | *AC025580.3* | *AL009176.1* | *AL354892.3* | *CPB2-AS1* | *MKLN1-AS* |
| *AC004528.1* | *AC034229.1* | *AL022311.1* | *AL355312.2* | *DDR1-DT* | *SLC16A1-AS1* |
| *AC004941.1* | *AC073320.2* | *AL022322.2* | *AL356495.1* | *FAM153CP* | *SNHG14* |
| *AC005616.1* | *AC080080.2* | *AL022323.2* | *AL356756.1* | *GEMIN7-AS1* | *UCKL1-AS1* |
| *AC007608.3* | *AC090559.1* | *AL022344.1* | *AL390860.1* | *HELLPAR* | *XACT* |
| *AC007613.1* | *AC092447.5* | *AL031595.3* | *AL590666.3* | *KDM5C-IT1* | *Z95114.1* |
| *AC009403.2* | *AC093028.1* | *AL078459.1* | *AL591163.1* | *LIN28B-AS1* | *Z95114.2* |
| *AC016705.2* | *AC109322.2* | *AL133410.1* | *AL592211.1* | *LINC00887* | |
| *AC016747.1* | *AC112504.1* | *AL137783.1* | *AL928654.1* | *LINC00963* | |
| *AC020659.1* | *AC132938.3* | *AL139317.5* | *AP001977.1* | *LINC01285* | |
| *AC020916.1* | *AC132938.7* | *AL157886.1* | *CCDC26* | *LINC02840* | |
| *AC020928.2* | *AL008628.1* | *AL353708.1* | *CELF2-DT* | *LSINCT5* | |
| *AL645728.1* | *AC025580.3* | *AL009176.1* | *AL354892.3* | *CPB2-AS1* | |

**Table S3.** Occurrence of the top 10 most prevalent lncRNA pG4 sequences in comparison to all pG4 motifs identified in lncRNA. Score = G4-iM Grinder Score. SINE = short interspersed nuclear element, LINE = long interspersed nuclear element, LTR = long terminal repeat.

| Top 10 most prevalent lncRNA G4 sequences | | | | | |
|---|---|---|---|---|---|
| Sequence | # of occurrences | % in repeats | Repeat Class(es) | # unique transcripts | # unique genes |
| GGGAGGCUGAGGCAGGAG | 560 | 54.1 | SINE, LINE, LTR | 508 | 358 |
| GGAGGCUGAGGCAGGAG | 365 | 57.6 | SINE, LINE, LTR | 344 | 233 |
| GGGAGGCUGAGGUGGG | 145 | 29.7 | SINE | 140 | 105 |
| GGGCGUGGUGGCGGG | 91 | 46.3 | SINE | 90 | 67 |
| GGGAGGCCGAGGCGGG | 81 | 33.8 | SINE | 80 | 65 |
| GGGAGGCUGAGGCGGG | 67 | 34.4 | SINE | 67 | 61 |
| GGAGGCUGAGGUGGG | 66 | 34.6 | SINE | 65 | 52 |
| GGGAGGCCGAGGUGGG | 86 | 20.0 | SINE | 86 | 45 |
| GAGGCUGAGGCAGGAG | 69 | 55.0 | SINE, LTR | 68 | 40 |
| GGGAGGCUGAGGCAGGAGAAU GGCGUGAACCCGGGAGGCG | 48 | 36.7 | SINE | 48 | 30 |
| | | | | | |
| All lncRNA pG4s (Score ≥ 20, MinNRuns ≥ 3) | 82,284 | 20.1 | - | 28,313 | 11,262 |
| All lncRNA pG4s (Score ≥ 40, MinNRuns ≥ 4) | 6,645 | 34.2 | - | 5,023 | 2,964 |
| All mRNA pG4s (Score ≥ 40, MinNRuns ≥ 4) | 29,049 | 25.8 | - | 20,046 | 7,912 |