Supporting Information

Thymine DNA Glycosylase is an RNA-Binding Protein with High Selectivity for G-Rich Sequences

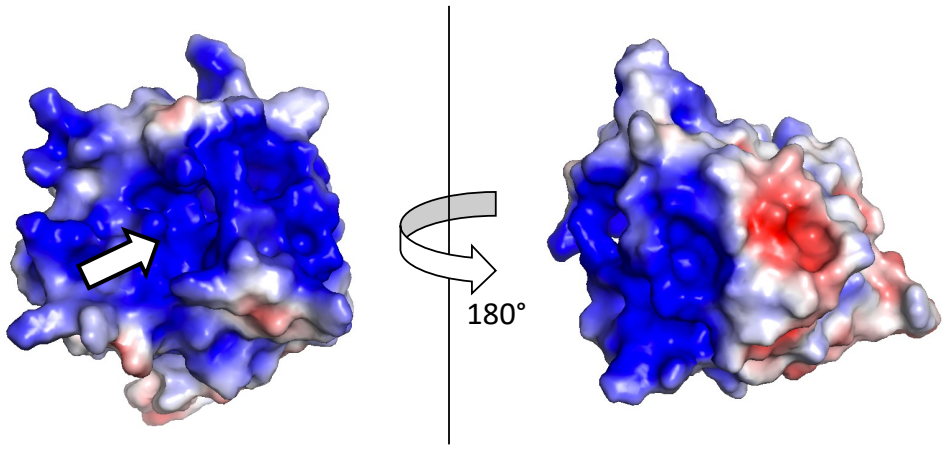
Lauren A. McGregor[1], Baiyu Zhu[1], Allison M. Goetz[1], and Jonathan T. Sczepanski[1,*]

[1]Department of Chemistry, Texas A&M University, College Station, Texas 77843

S1. Supplementary Figures.

Figure S1

a



b

NTD	MEAENAGSYSLQQAQAFYTFPFQQLMAEAPNMAVVNEQQMPEEVPAPAPAQEPVQEAPKGR KRKPRTTEPKQPVEPKKPVESKKSGKSAKSKEKQEKITDTFKVKRKVDR
CAT	FNGVSEAELLTKTLPDILTFNLDIVIIGINPGLMAAYKGHHYPGPGNHFWKCLFMSGLSEVQLNH MDDHTLPGKYGIGFTNMVERTTPGSKDLSSKEFREGGRILVQKLQKYQPRIAVFNGKCIYEIFSK EVFGVKVKNLEFGLQPHKIPDTETLCYVMPSSSARCAQFPRAQDKVHYYIKLKDLRDQLKGIER NMD
CTD	VQEVQYTFDLQLAQEDAKKMAVKEEKYDPGYEAAYGGAYGENPCSSEPCGFSSNGLIESVELR GESAFSGIPNGQWMTQSFTDQIPSFSNHCGTQEQEEESHA

Figure S1. (a) Electrostatic potential surface of TDG (catalytic domain, residues 111-308) determined from PDB ID:3UFJ.[1] Negative potentials are depicted by the red color and positive potentials by the blue color. The arrow indicates the active site. (b) Amino acid sequence of TDG, separated by the N-terminal domain (NTD), catalytic domain (CAT), and C-terminal domain (CTD). Cationic and polar residues are shown in blue and green text, respectively, for the disordered NTD and CTD.

Figure S2

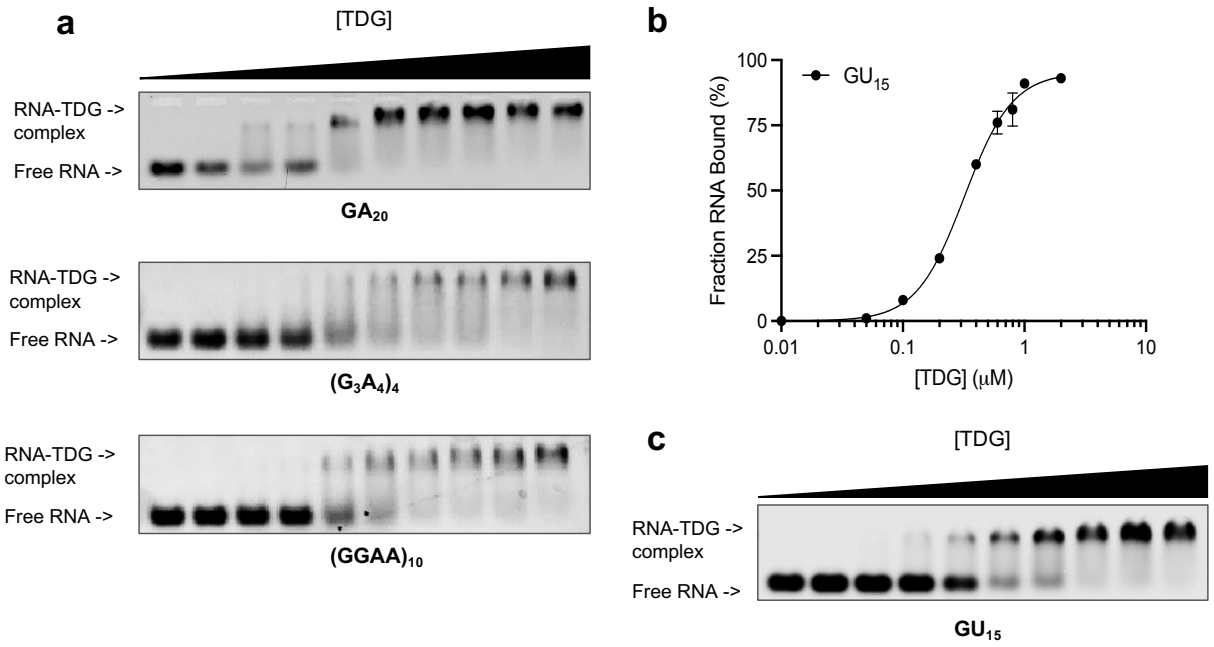


Figure S2. TDG binds preferentially to G-rich RNA *in vitro*. (a) Representative EMSA data for G-rich RNAs containing different arrangements of Gs and As binding to TDG (0 – 5 μM). (b) Saturation plots for binding of TDG to GU_{15} RNA sequence. Data are mean ± S.D. (n = 3). (c) Representative EMSA data for GU_{15} RNA sequence binding to TDG (0 – 2 μM). Uncropped gel images are presented in Figure S13.

Figure S3

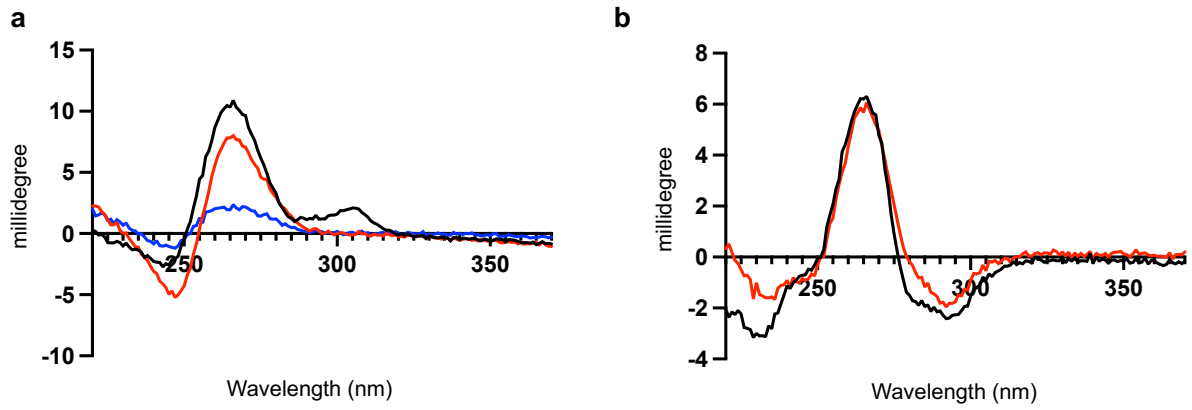


Figure S3. (a) Circular dichroism (CD) spectra of 5 μM $(GGAA)_{10}$ (black), $(G_3A_4)_4$ (red) and $(GA)_{20}$ (blue) RNAs in a buffer containing 37.5 mM NaCl, 12.5 mM KCl, 10 mM HEPES (pH 7.8), 2.5 mM BME, and 5% glycerol. The CD spectra of $(GGAA)_{10}$ and $(G_3A_4)_4$ are consistent with a parallel G4 structure in the presence of K^+ as evident by a positive band at ~265 nm and a negative band at ~240 nm. These features are substantially reduced in $(GA)_{20}$, indicating little to no G4 formation. (b) Circular dichroism (CD) spectra of 5 μM CG_{HP} (black) and MUT_{HP} (red) RNAs in a buffer containing 37.5 mM NaCl, 12.5 mM KCl, 10 mM HEPES (pH 7.8), 2.5 mM BME, and 5% glycerol. The CD spectrum of confirms the formation of a hairpin structure as evident by a negative band at ~290 nm and a positive band at ~265 nm.

Figure S4

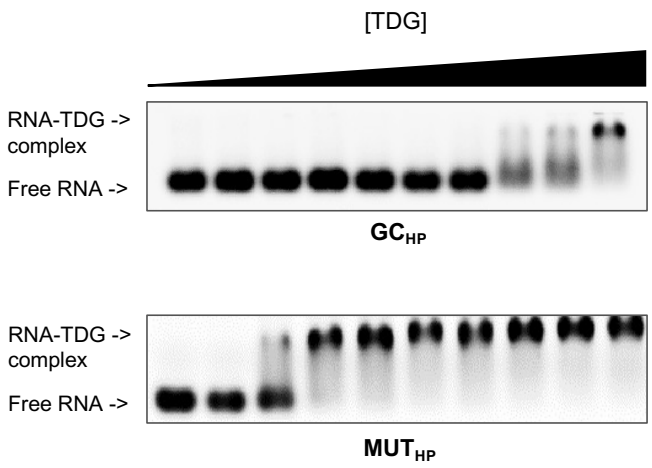


Figure S4. Representative EMSA data for CG_{HP} and MUT_{HP} RNAs binding to TDG (0 – 2 μM) *in vitro*. Uncropped gel images are presented in Figure S14.

Figure S5

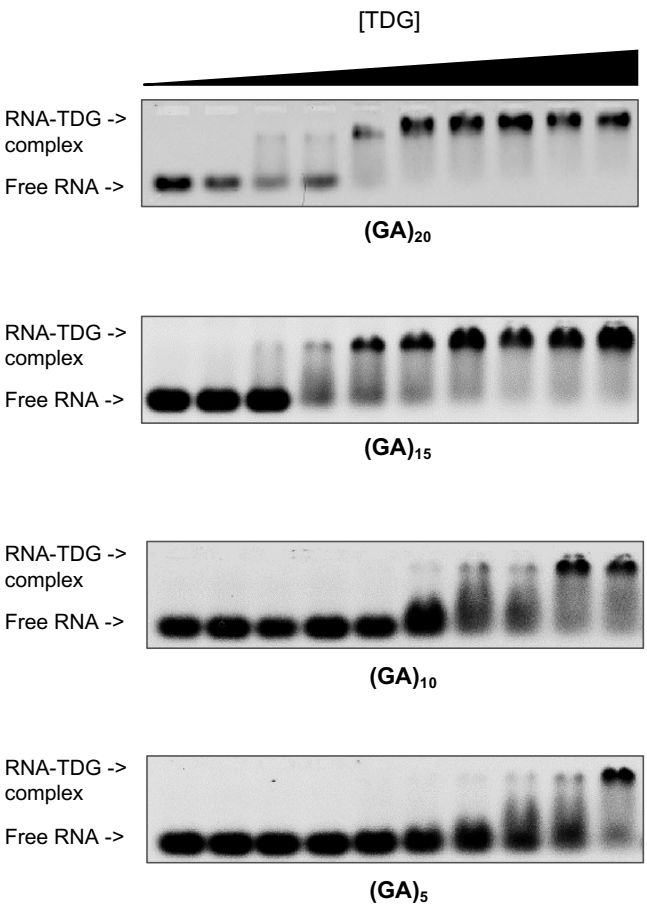


Figure S5. Representative EMSA data for $(GA)_{20}$ RNA and its truncations binding to TDG (0 – 2 μM). Uncropped gel images are presented in Figure S15. The image for $(GA)_{20}$ was reused from Figure S2a and was positioned here to allow for the convenient visual comparison of $(GA)_{20}$ to its truncated variants.

Figure S6

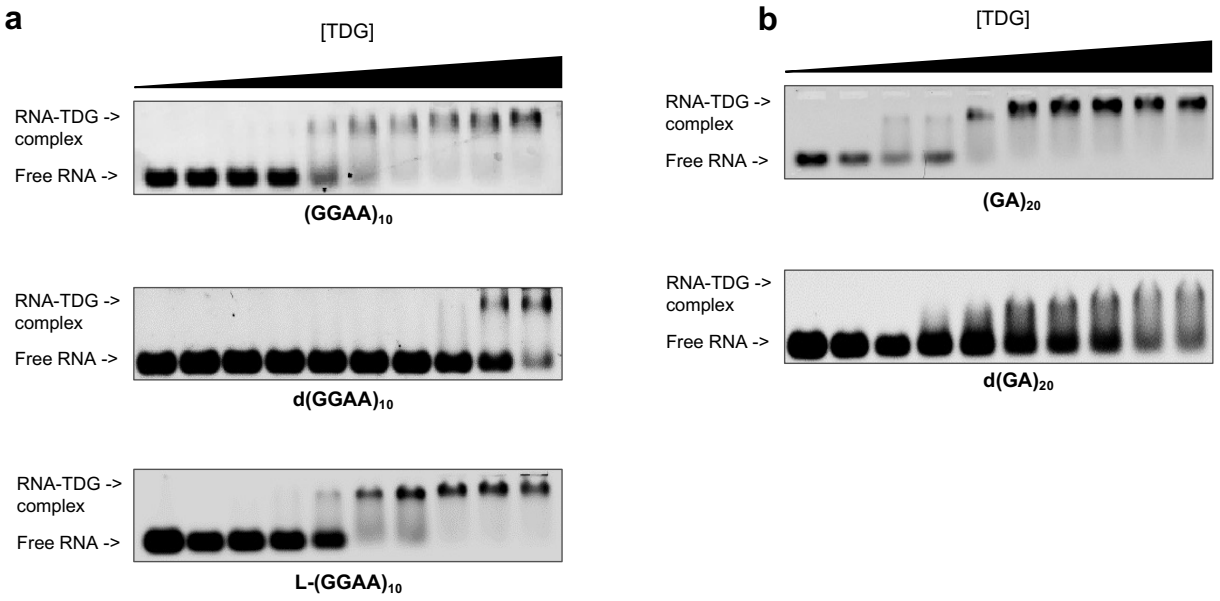
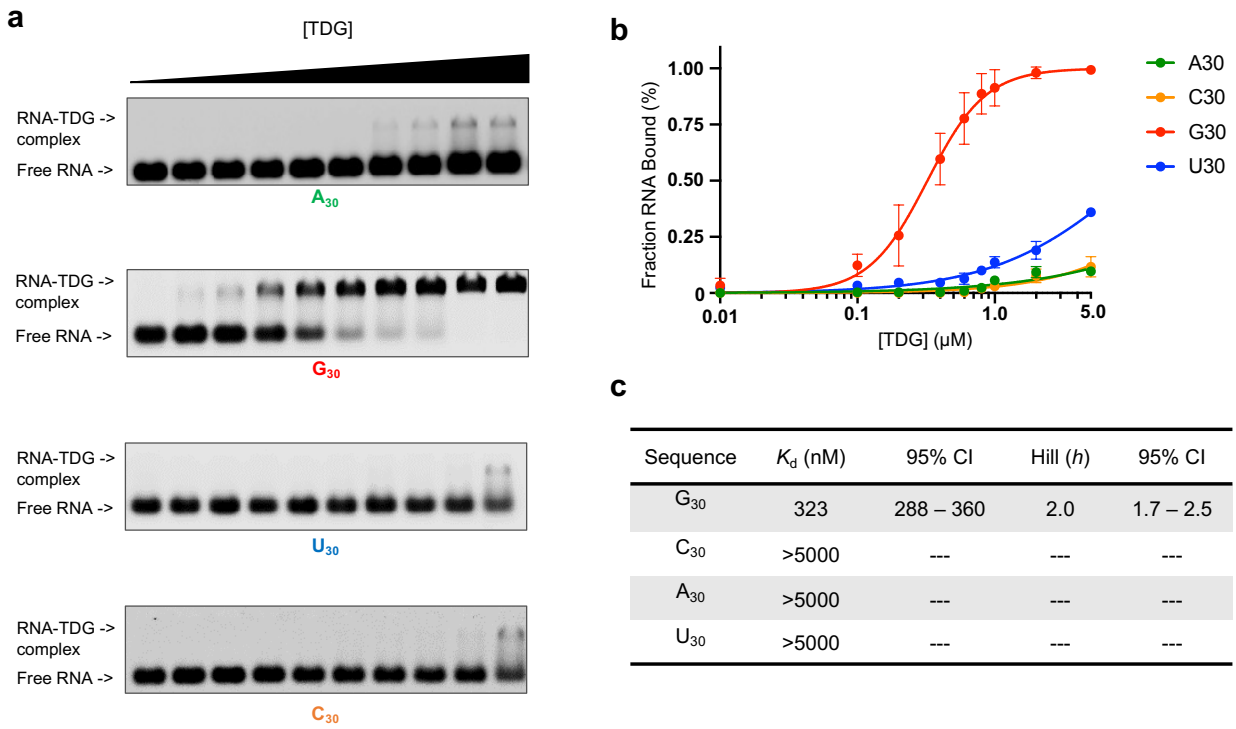


Figure S6. Representative EMSA data for $(GGAA)_{10}$ (a) and $(GA)_{20}$ (b) series of oligonucleotides binding to TDG *in vitro*. Uncropped gel images are presented in Figure S16. The image for $(GGAA)_{10}$ in panel a was reused from Figure S2a and was positioned here to allow for the convenient visual comparison of $(GGAA)_{10}$ to its DNA and L-DNA counterparts. The image for $(GA)_{20}$ in panel b was reused from Figure S2a and was positioned here to allow for the convenient visual comparison of $(GA)_{20}$ to its DNA counterpart $d(GA)_{20}$.

Figure S7

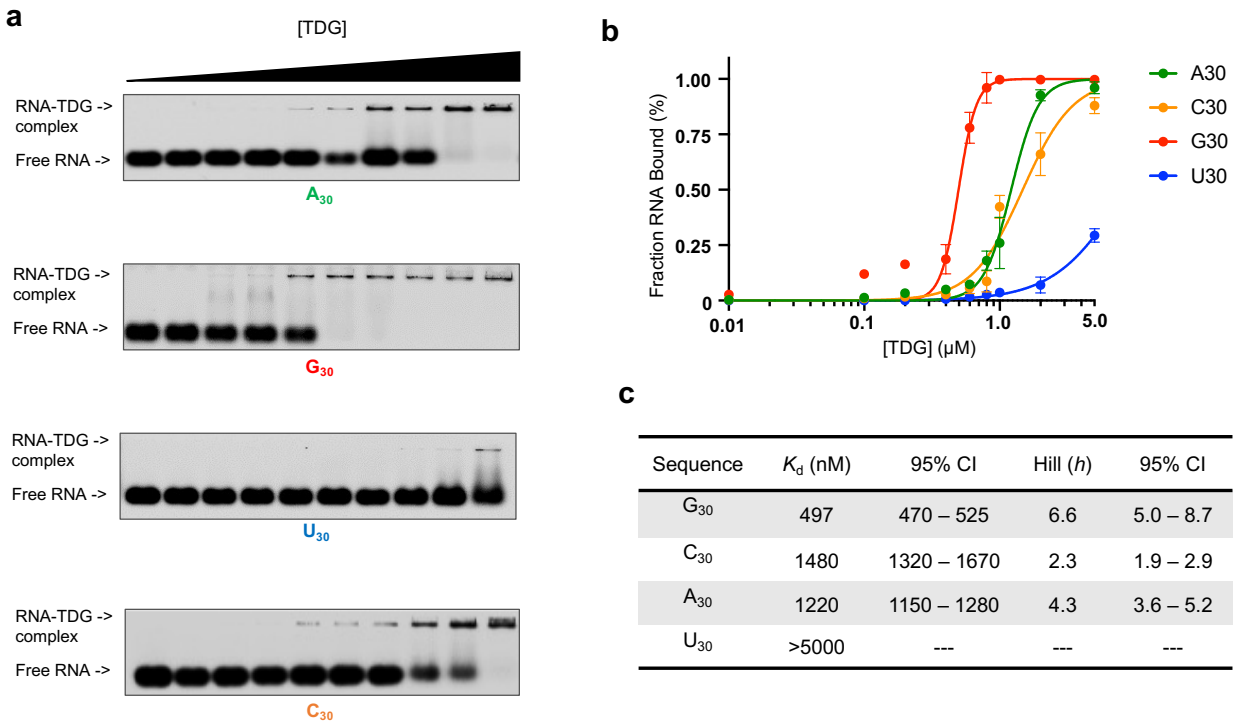


c

Sequence	K_d (nM)	95% CI	Hill (*h*)	95% CI
G_{30}	323	288 – 360	2.0	1.7 – 2.5
C_{30}	>5000	---	---	---
A_{30}	>5000	---	---	---
U_{30}	>5000	---	---	---

Figure S7. The influence of TDG's NTD deletion ($TDG_{\Delta N}$) on RNA binding. (a) Representative EMSAs for homopolymeric RNA sequences binding to $TDG_{\Delta N}$. Uncropped gel images are presented in Figure S17. (b) Saturation plots for binding of $TDG_{\Delta N}$ binding to homopolymeric RNAs. Data are mean ± S.D. (n = 3). (c) Equilibrium dissociation constants and Hill coefficients (*h*) for $TDG_{\Delta N}$ binding to various RNAs. 95% confidence interval (95% CI).

Figure S8

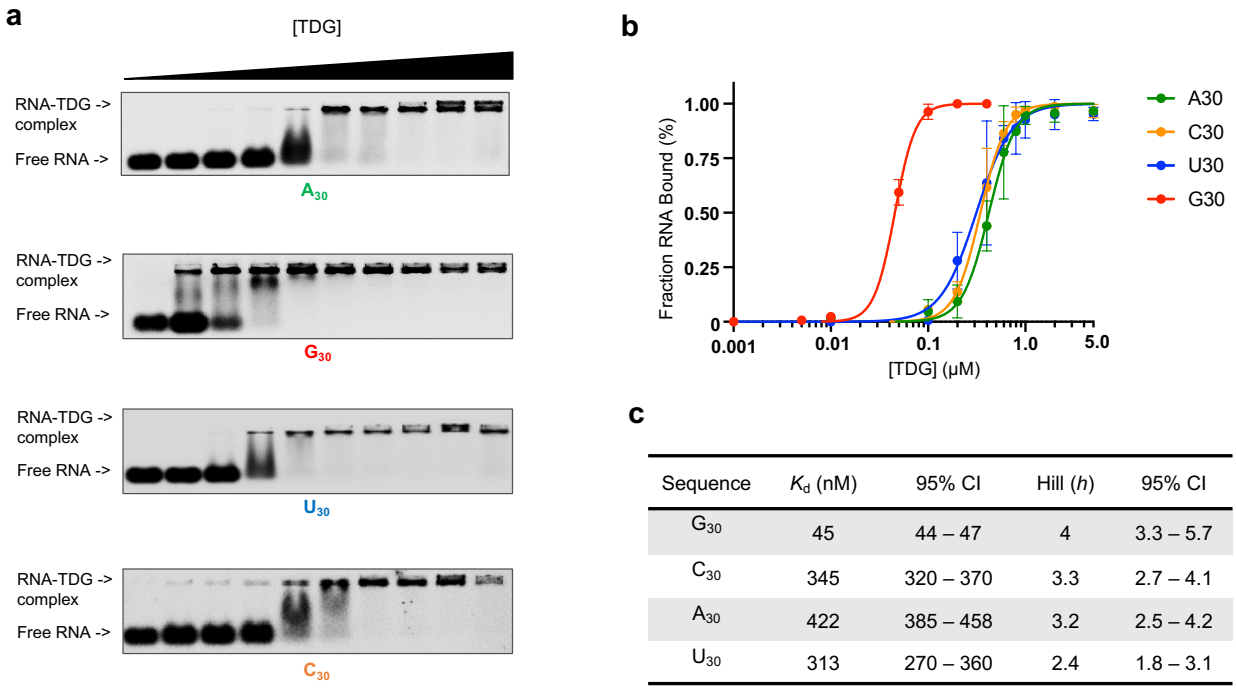


c

Sequence	K_d (nM)	95% CI	Hill (h)	95% CI
G_{30}	497	470 – 525	6.6	5.0 – 8.7
C_{30}	1480	1320 – 1670	2.3	1.9 – 2.9
A_{30}	1220	1150 – 1280	4.3	3.6 – 5.2
U_{30}	>5000	---	---	---

Figure S8. The influence of TDG's catalytic domain (TDG_{CAT}) on RNA binding. (a) Representative EMSAs for homopolymeric RNA sequences binding to TDG_{CAT}. Uncropped gel images are presented in Figure S18. (b) Saturation plots for binding of TDG_{CAT} binding to homopolymeric RNAs. Data are mean ± S.D. (n = 3). (c) Equilibrium dissociation constants and Hill coefficients (h) for TDG_{CAT} binding to various RNAs. 95% confidence interval (95% CI).

Figure S9



c

Sequence	K_d (nM)	95% CI	Hill (*h*)	95% CI
G_{30}	45	44 – 47	4	3.3 – 5.7
C_{30}	345	320 – 370	3.3	2.7 – 4.1
A_{30}	422	385 – 458	3.2	2.5 – 4.2
U_{30}	313	270 – 360	2.4	1.8 – 3.1

Figure S9. The influence of TDG's CTD deletion ($TDG_{\Delta C}$) on RNA binding. (a) Representative EMSAs for homopolymeric RNA sequences binding to $TDG_{\Delta C}$. Uncropped gel images are presented in Figure S19. (b) Saturation plots for binding of $TDG_{\Delta C}$ binding to homopolymeric RNAs. Data are mean ± S.D. (n = 3). (c) Equilibrium dissociation constants and Hill coefficients (*h*) for $TDG_{\Delta C}$ binding to various RNAs. 95% confidence interval (95% CI).

Figure S10

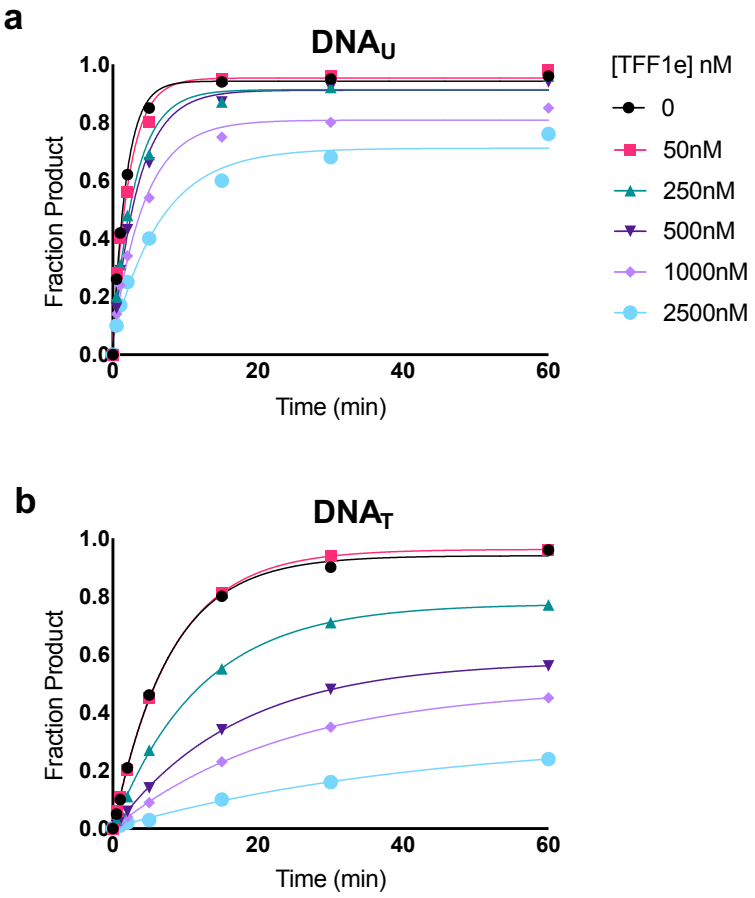


Figure S10. Excision of DNA_U (a) and DNA_T (b) mismatched substrates by TDG is inhibited by native TFF1e RNA. (a) TDG-mediated excision of G•U mismatch is inhibited by TFF1e RNA as concentration of RNA increased. (b) TDG-mediated excision of G•T mismatch is inhibited by TFF1e RNA, drastically. For each reaction, the DNA substrate (100 nM) was mixed with the indicated concentration of G_{30} RNA followed by the addition of TDG (200 nM).

Figure S11

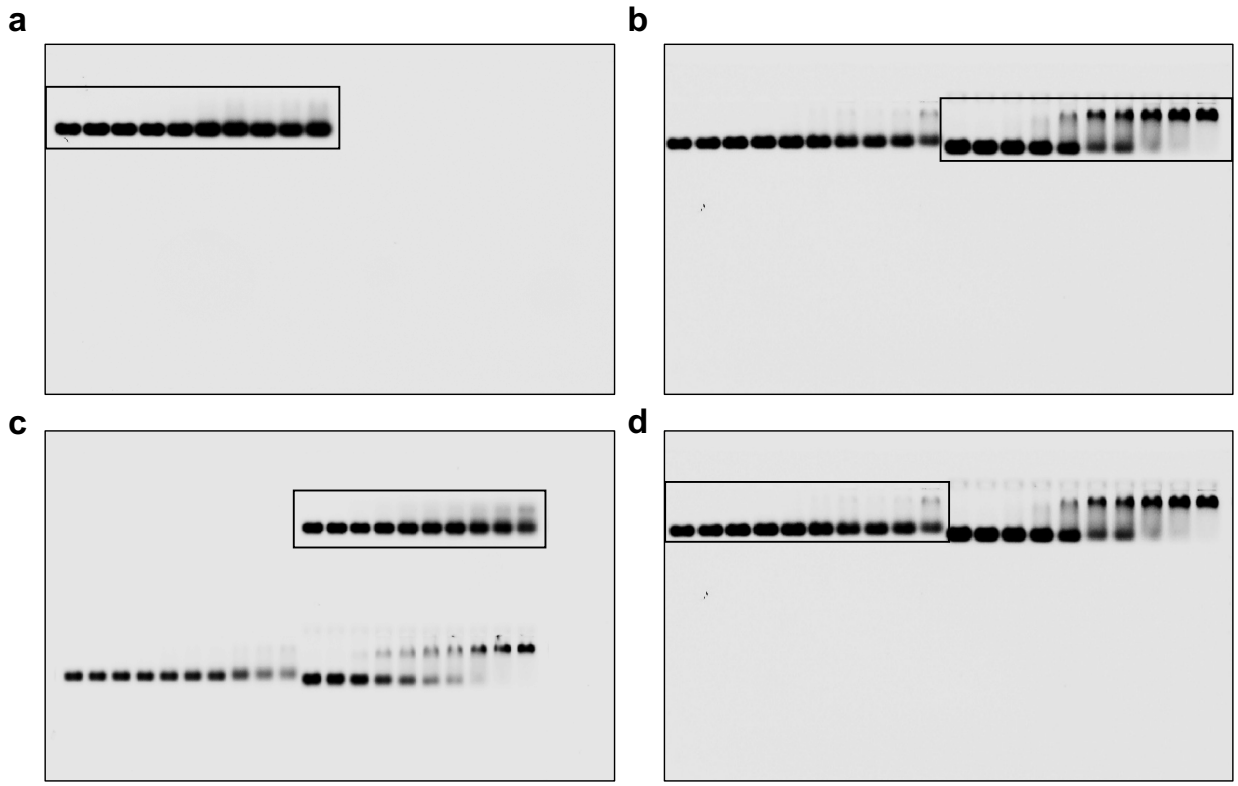


Figure S11. Uncropped gel images for main text Figure 1a. (a) A_{30}. (b) G_{30}. (c) U_{30}. (d) C_{30}. Frame indicates the cropped region shown in Figure 1a. The gel in panel c has unrelated experiments originating from wells at the midpoint of the gel.

Figure S12

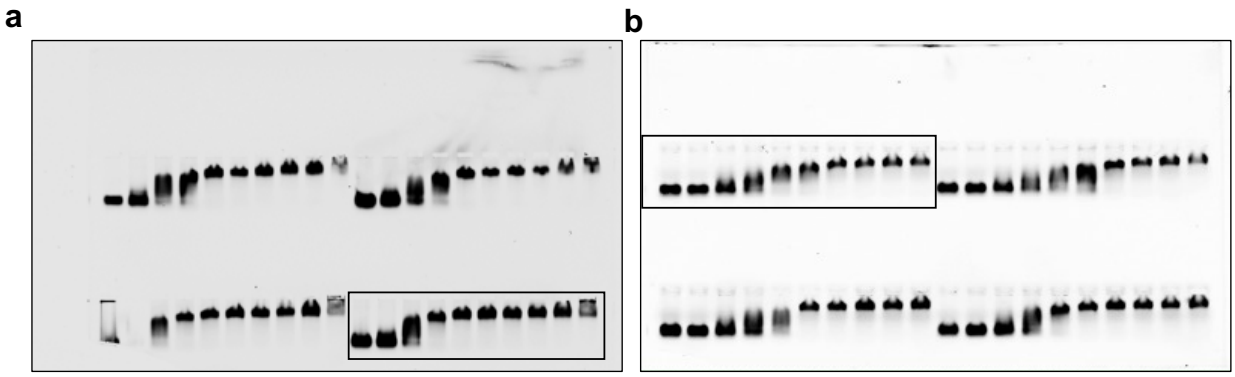


Figure S12. Uncropped gel images for main text Figure 4a. (a) HOTAIR. (b) TFF1e. Frame indicates the cropped region shown in Figure 4a. The gel in panel a has unrelated experiments originating from wells at the top of the gel. The gel in panel b has unrelated experiments originating from wells at the midpoint of the gel.

Figure S13

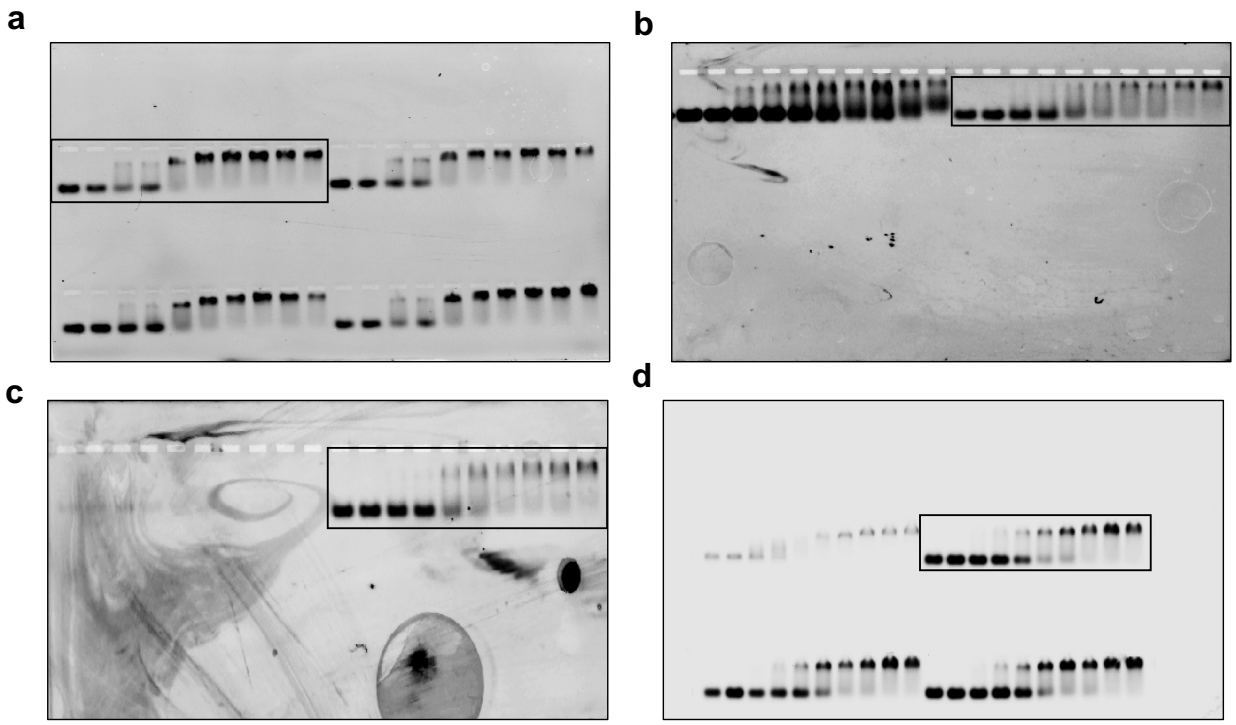


Figure S13. Uncropped gel images for Figure S2a,c. (a) GA_{20}. (b) $(G_3A_4)_4$. (c) $(GGAA)_{10}$. (d) GU_{15}. Frame indicates the cropped region shown in Figure S2a,c. The gels in panels a and d have unrelated experiments originating from wells at the midpoint of the gel.

Figure S14

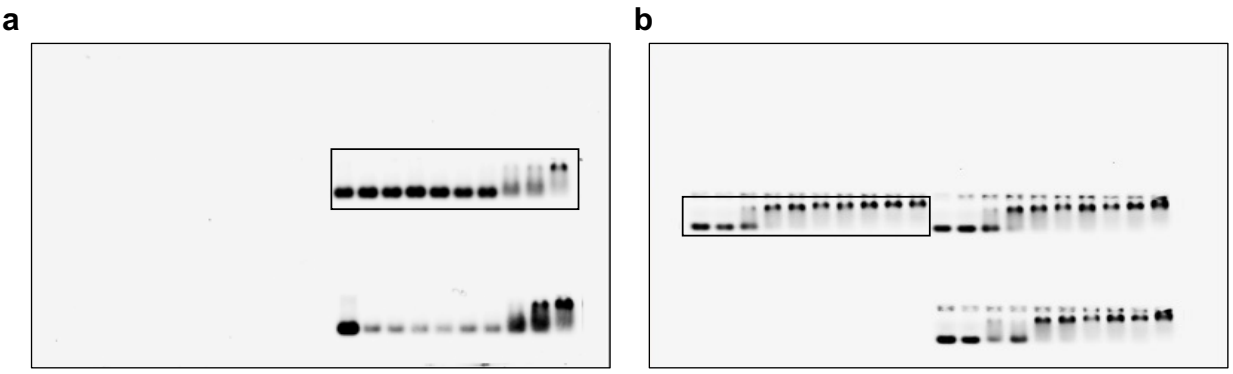


Figure S14. Uncropped gel images for Figure S4. (a) GC_{HP}. (b) MUT_{HP}. Frame indicates the cropped region shown in Figure S4. Gels have unrelated experiments originating from wells at the midpoint of the gel.

Figure S15

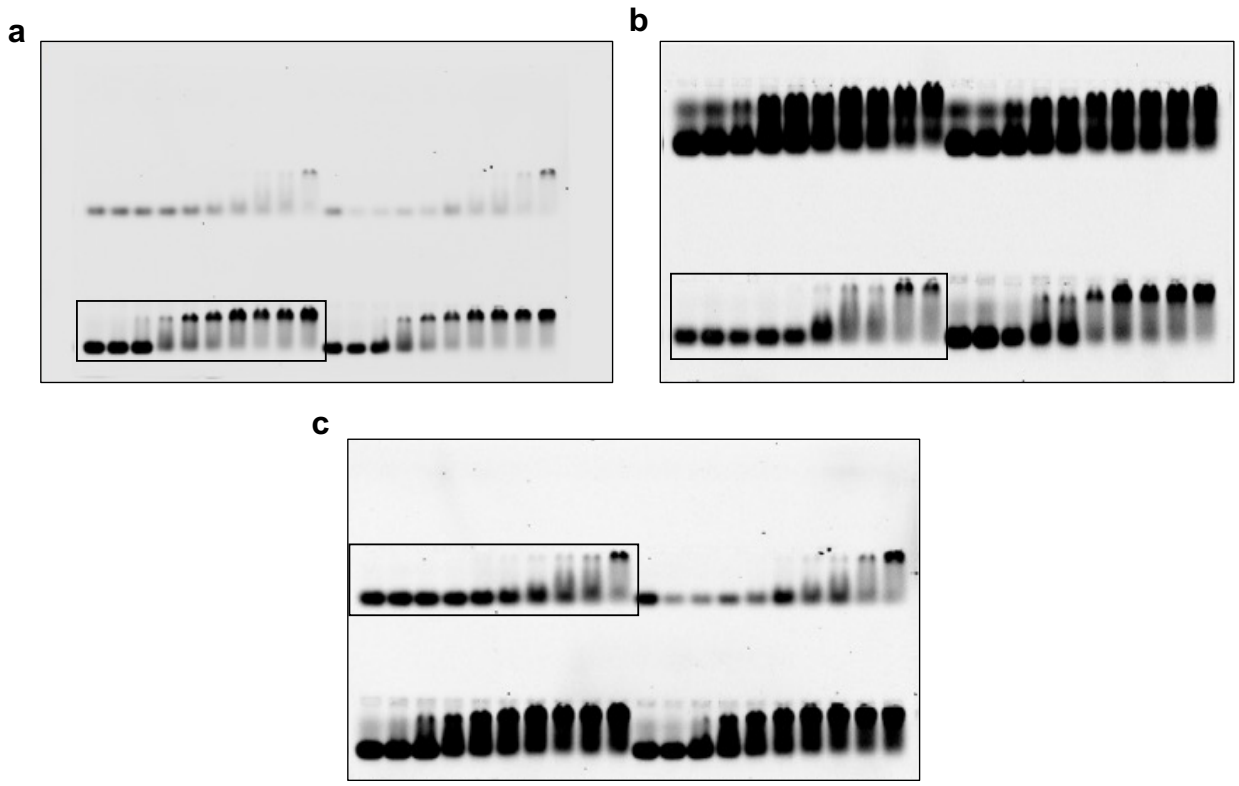


Figure S15. Uncropped gel images for Figure S5. (a) $(GA)_{15}$. (b) $(GA)_{10}$. (c) $(GA)_5$. Frame indicates the cropped region shown in Figure S5. The gel in panels c has unrelated experiments originating from wells at the midpoint of the gel. The gels in panels a and b have unrelated experiments originating from wells at the top of the gel.

Figure S16

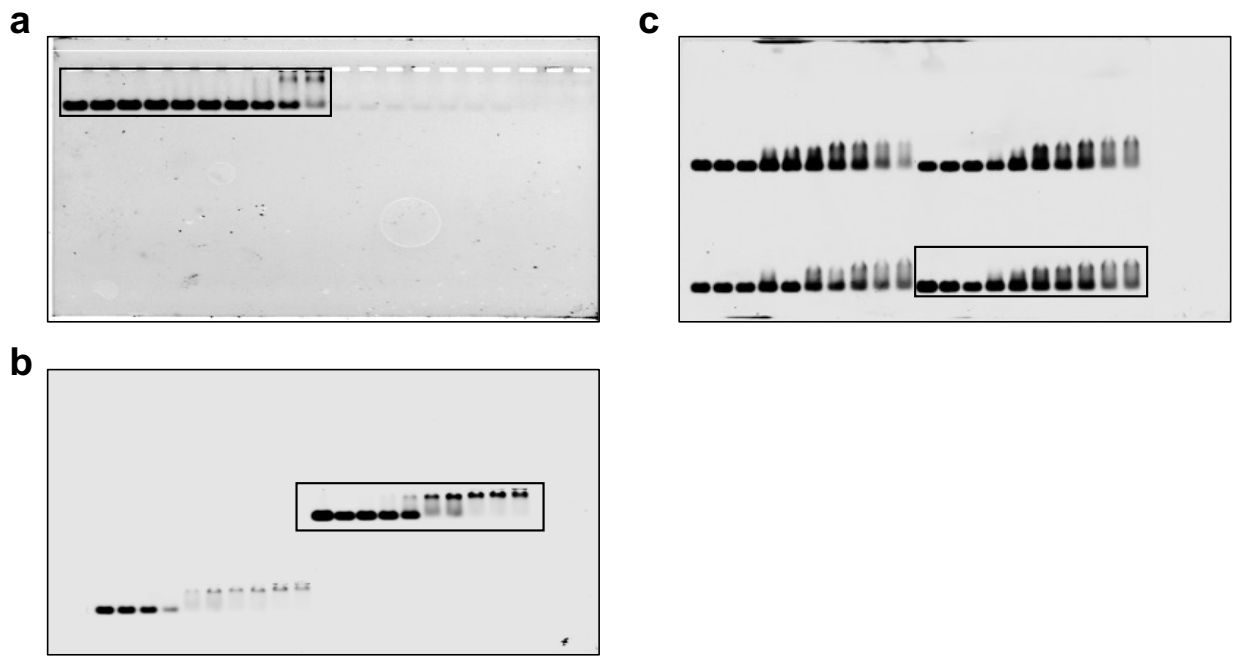


Figure S16. Uncropped gel images for Figure S6. (a) d$(GGAA)_{10}$. (b) L-$(GGAA)_{10}$. (c) d$(GA)_{20}$. Frame indicates the cropped region shown in Figure S6. The gel in panels b has unrelated experiments originating from wells at the midpoint of the gel. The gel in panel c has unrelated experiments originating from wells at the top of the gel.

Figure S17

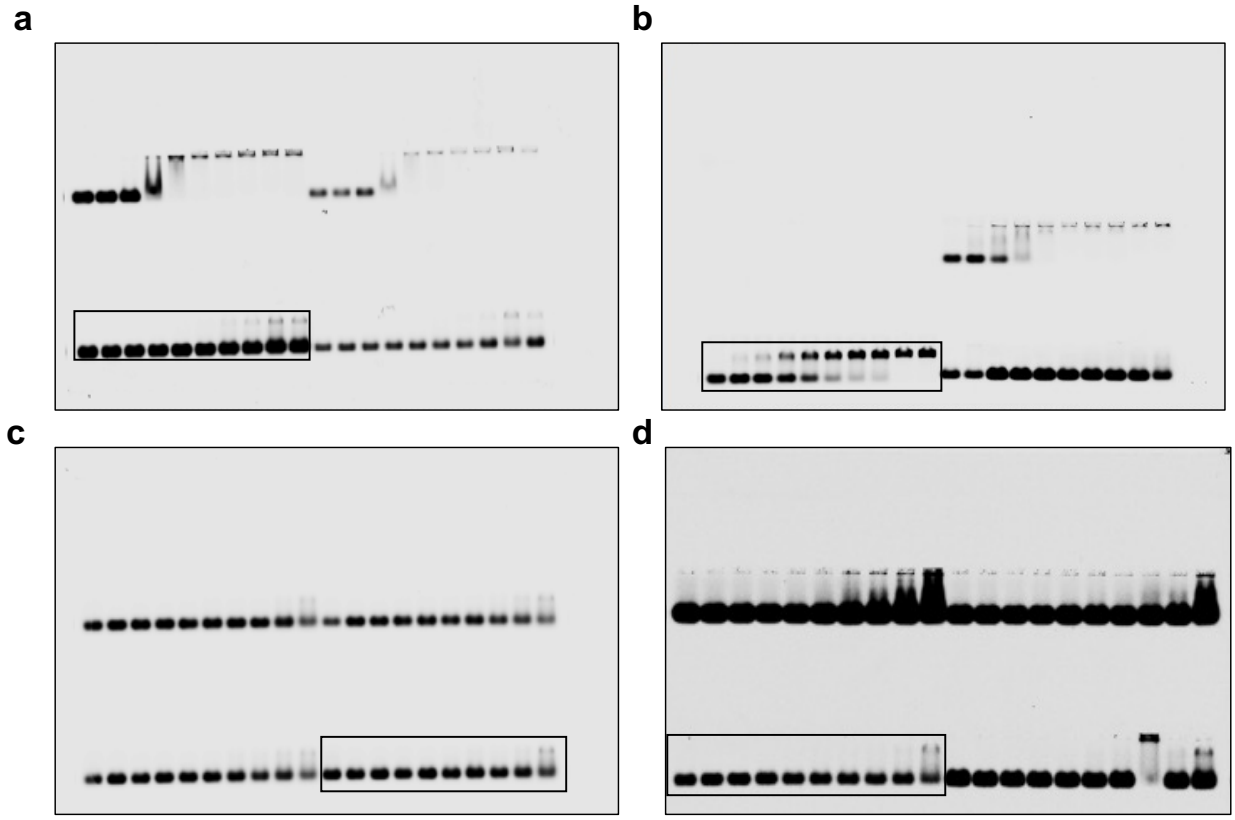


Figure S17. Uncropped gel images for Figure S7a. (a) A_{30}. (b) G_{30}. (c) U_{30}. (d) C_{30}. Frame indicates the cropped region shown in Figure S7a. The gels have unrelated experiments originating from wells at the top of the gel.

Figure S18

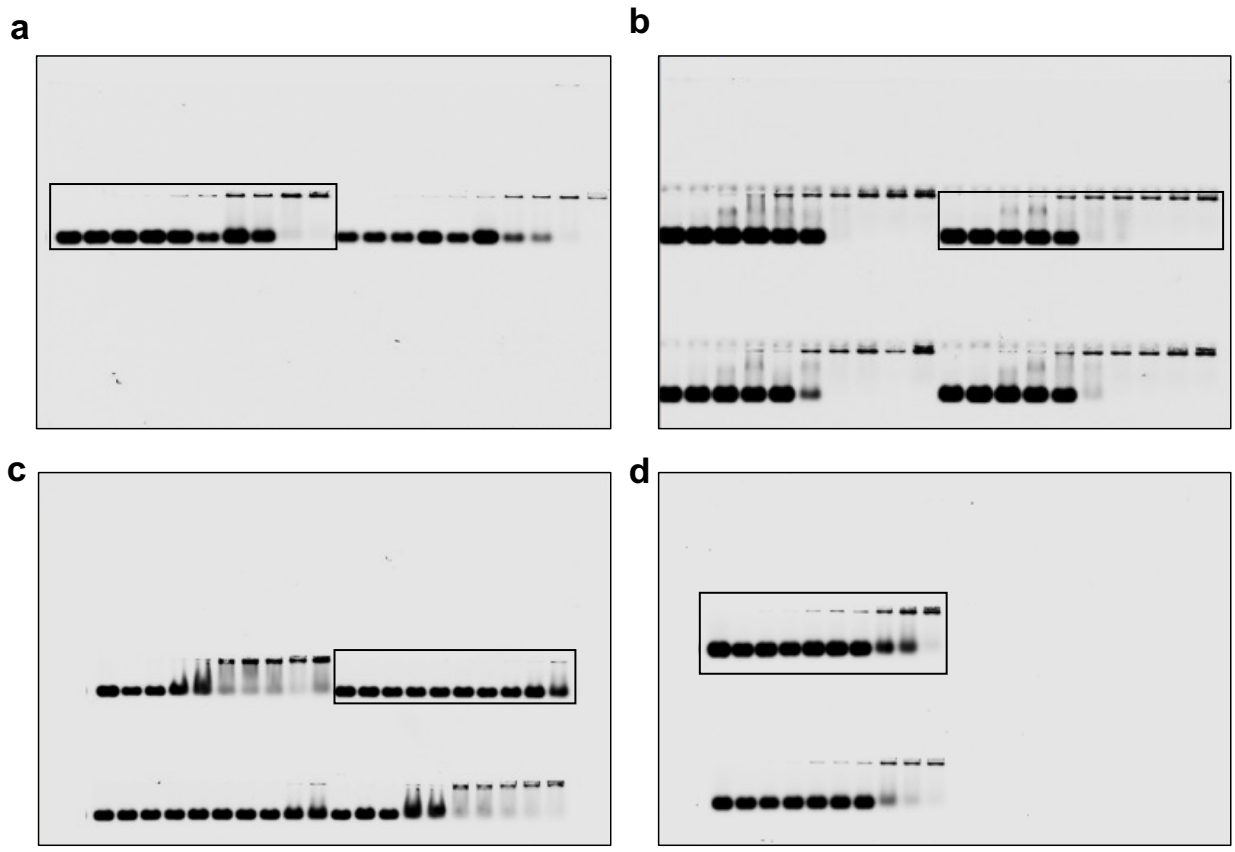


Figure S18. Uncropped gel images for Figure S8a. (a) A_{30}. (b) G_{30}. (c) U_{30}. (d) C_{30}. Frame indicates the cropped region shown in Figure S8a. The gel in panel b has unrelated experiments originating from wells at the top of the gel. The gels in panels c and d have unrelated experiments originating from wells at the midpoint of the gel.

Figure S19

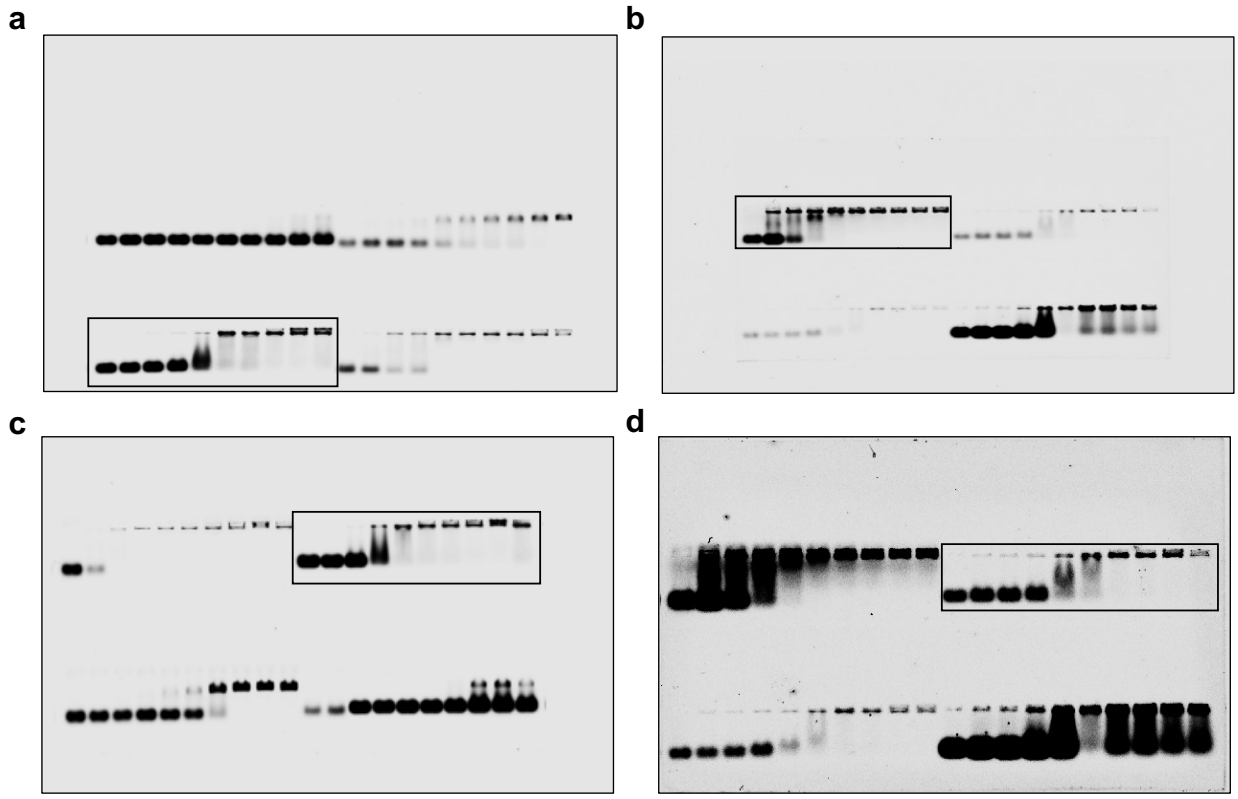


Figure S19. Uncropped gel images for Figure S9a. (a) A_{30}. (b) G_{30}. (c) U_{30}. (d) C_{30}. Frame indicates the cropped region shown in Figure S9a. The gel in panel a has unrelated experiments originating from wells at the top of the gel. The gels in panels b, c and d have unrelated experiments originating from wells at the midpoint of the gel.

Figure S20

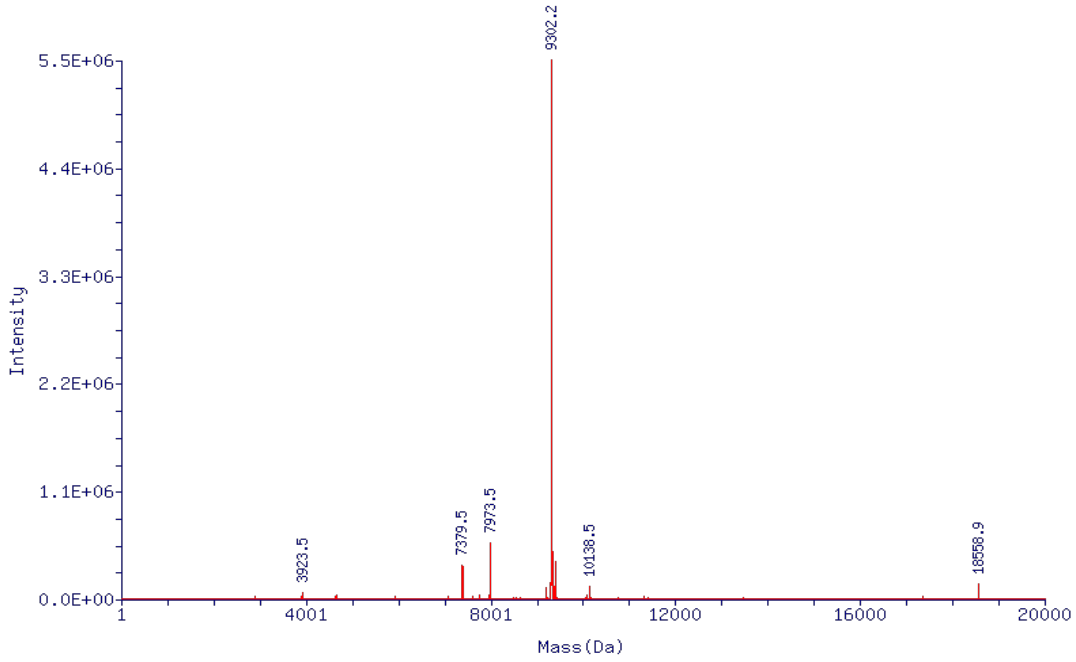


Figure S20. ESI-MS spectra of U_{30} prepared by solid-phase synthesis. Mass calculated: 9,302.3 Da; Mass found: 9,302.2 Da.

Figure S21

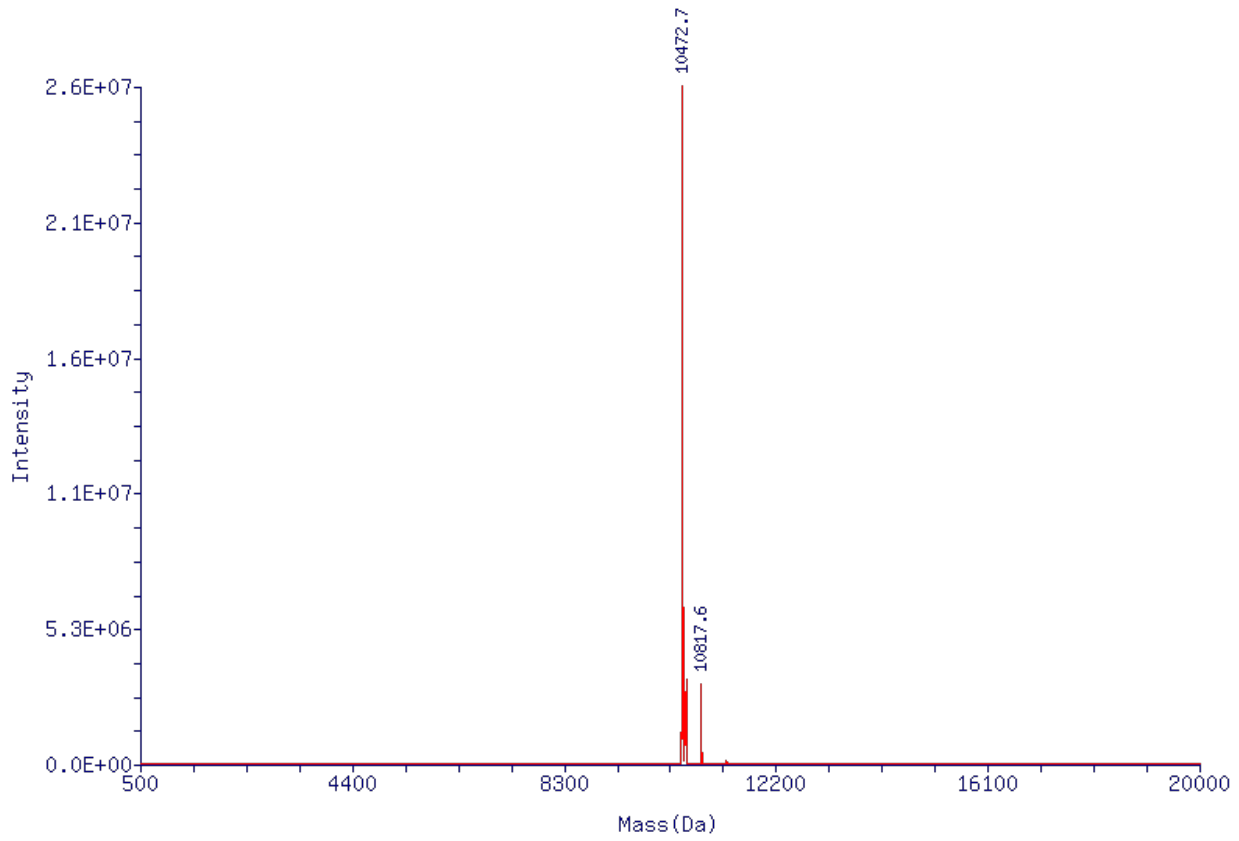


Figure S21. ESI-MS spectra of G_{30} prepared by solid-phase synthesis. Mass calculated: 10,473.4 Da; Mass found: 10,472.7 Da.

Figure S22

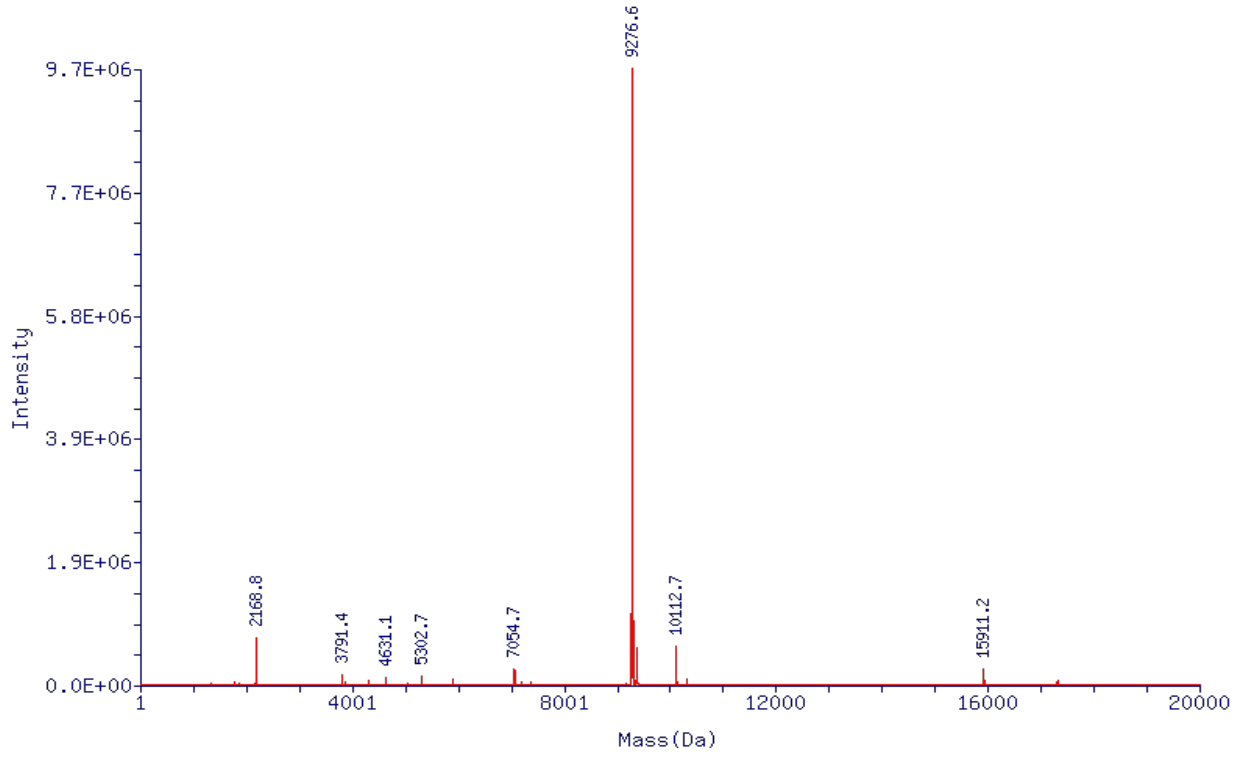


Figure S22. ESI-MS spectra of C_{30} prepared by solid-phase synthesis. Mass calculated: 9272.65 Da; Mass found: 9,276.6 Da.

Figure S23

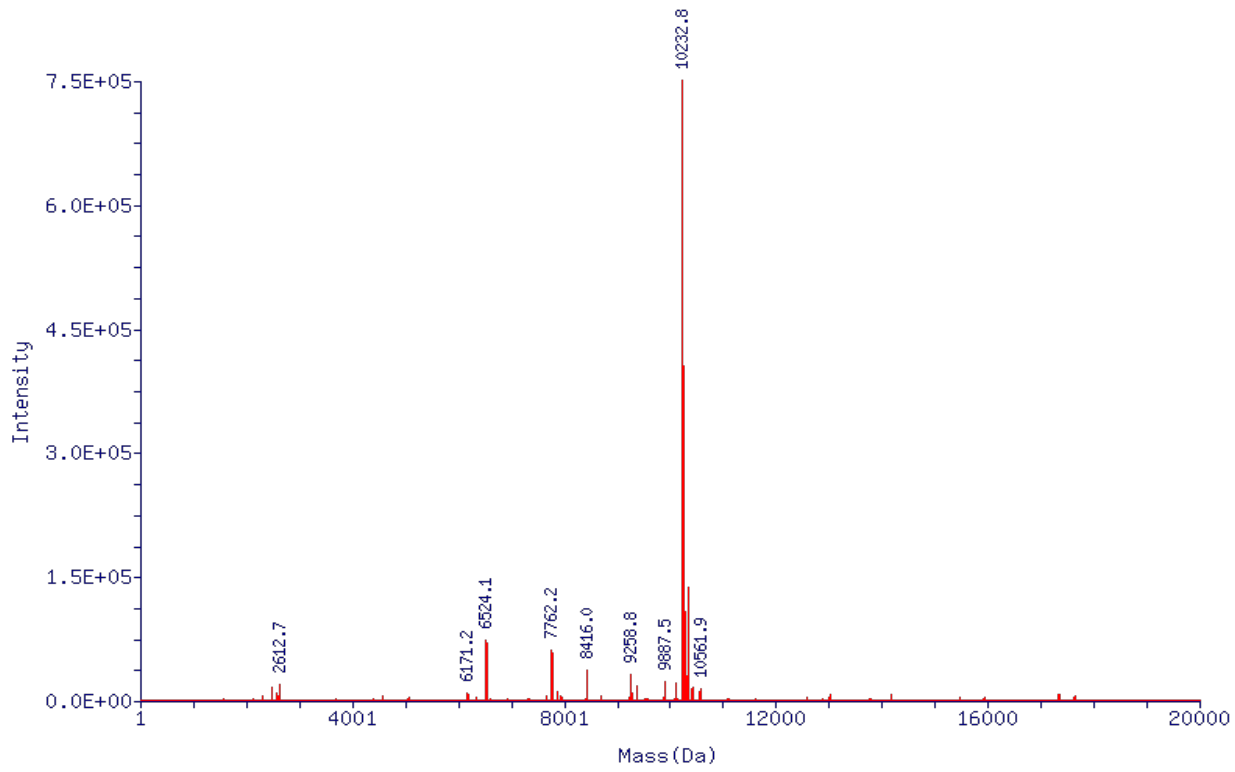


Figure S23. ESI-MS spectra of $(GA)_{15}$ prepared by solid-phase synthesis. Mass calculated: 10,233.45 Da; Mass found:10,232.8 Da.

Figure S24

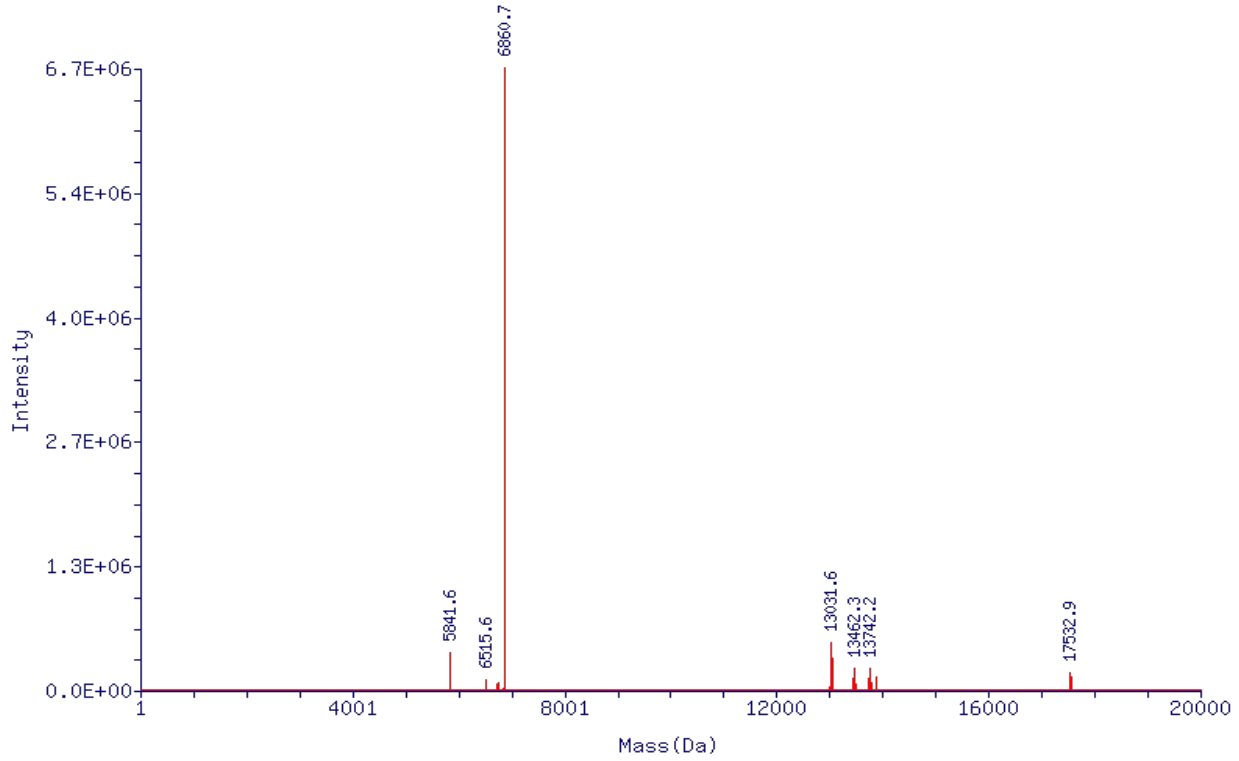


Figure S24. ESI-MS spectra of $(GA)_{10}$ prepared by solid-phase synthesis. Mass calculated: 6861.35 Da; Mass found: 6860.7 Da.

Figure S25

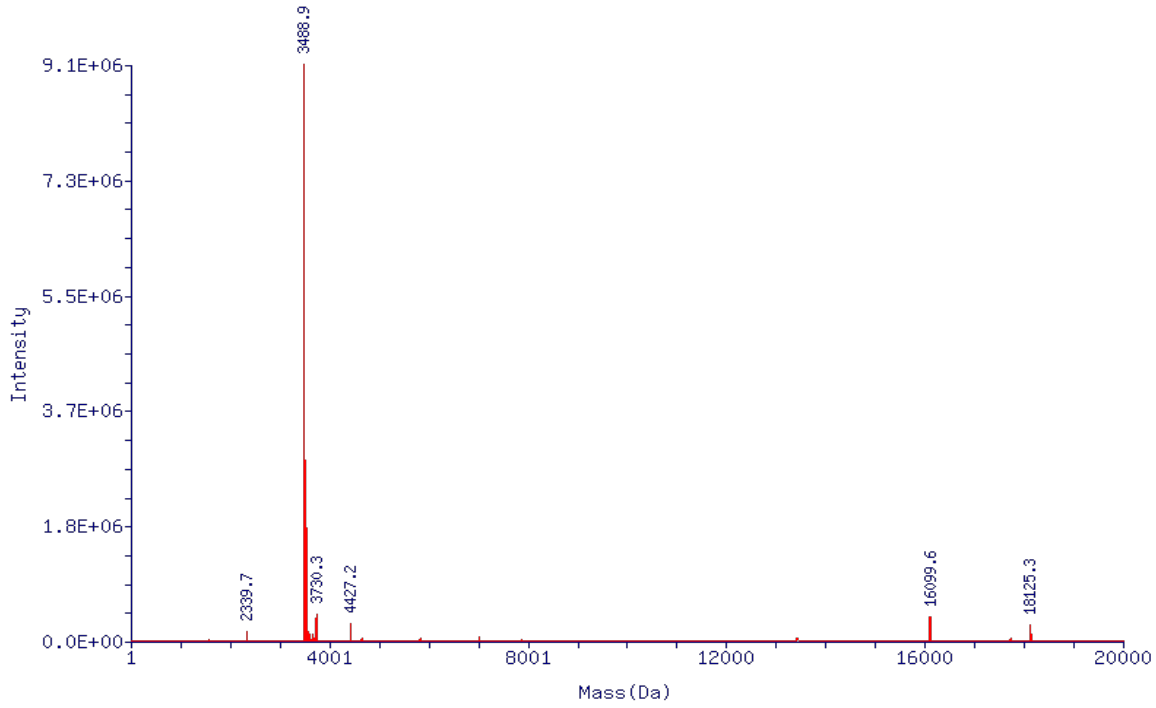


Figure S25. ESI-MS spectra of $(GA)_5$ prepared by solid-phase synthesis. Mass calculated: 3,489.25 Da; Mass found: 3,488.9 Da.

Figure S26

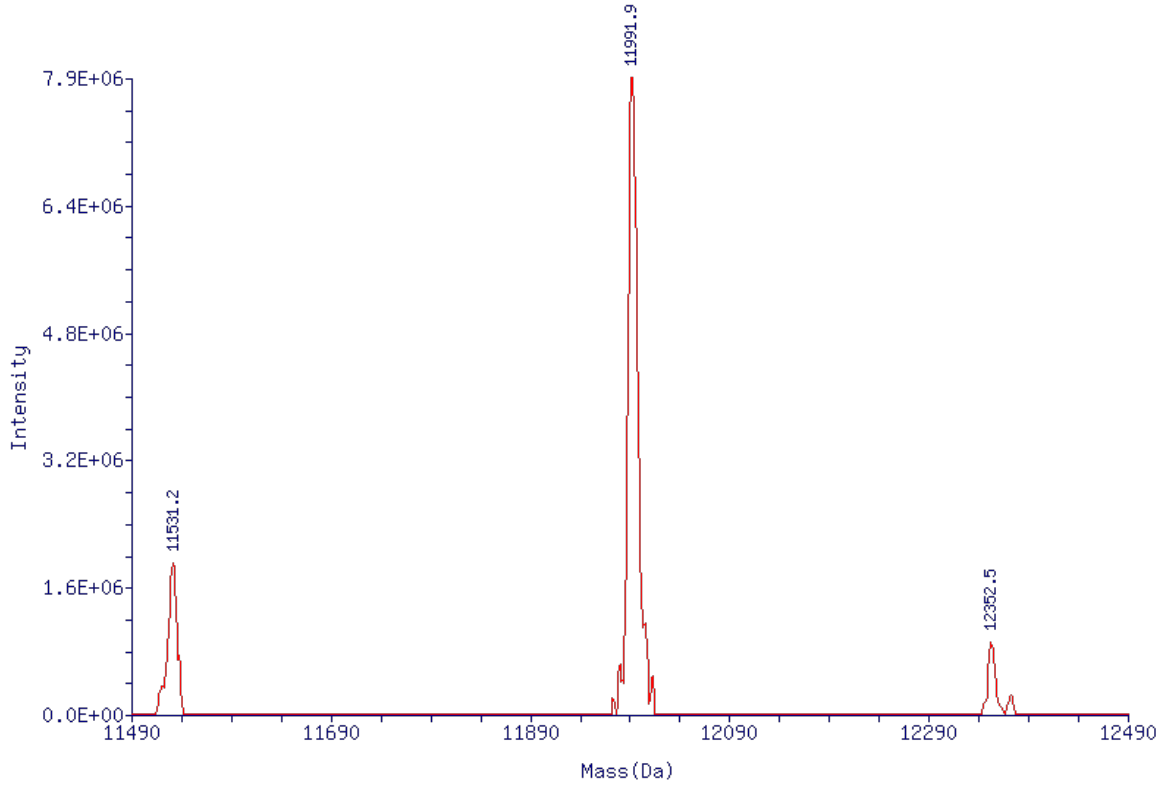


Figure S26. ESI-MS spectra of MUT_{HP} prepared by solid-phase synthesis. Mass calculated: 11992.35 Da; Mass found:11991.9 Da.

Figure S27

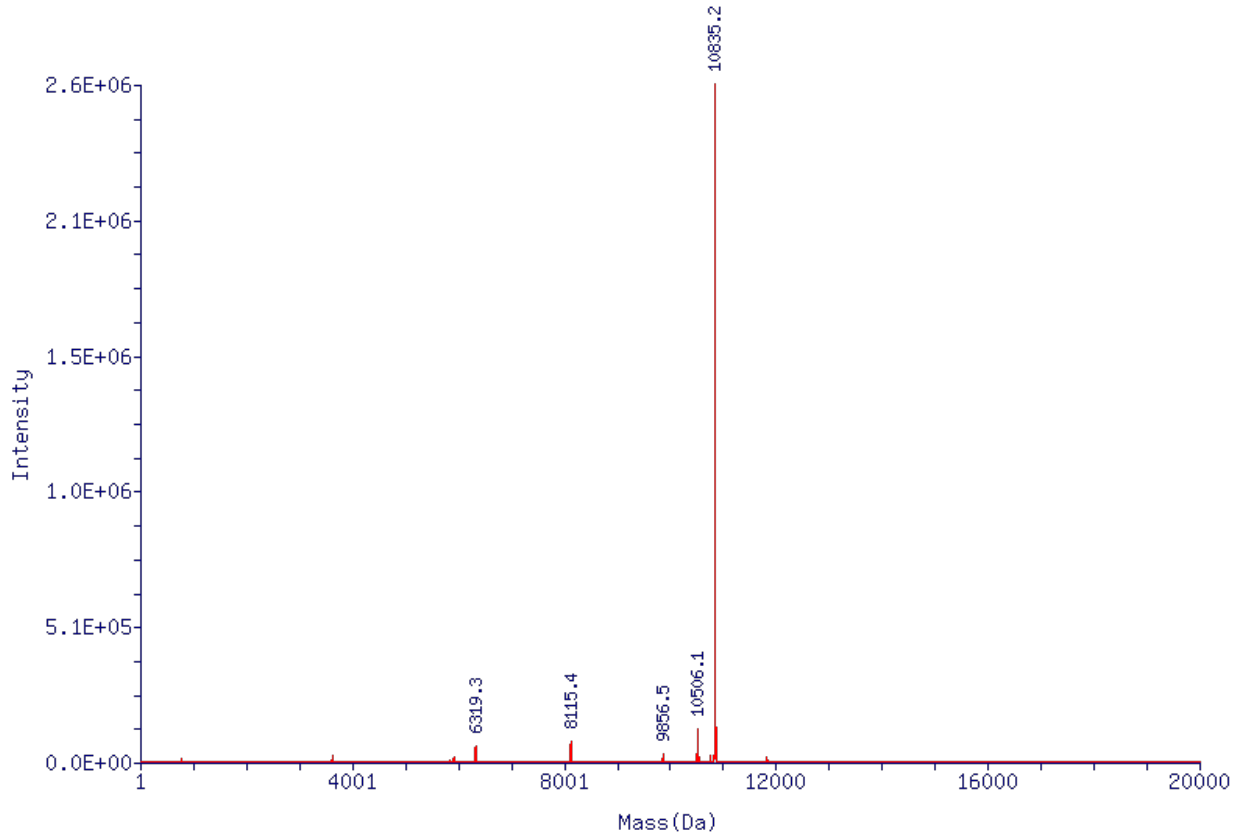


Figure S27. ESI-MS spectra of ^{F}U prepared by solid-phase synthesis. Mass calculated: 10,835.1 Da; Mass found: 10835.2 Da.

S2. Supplementary Tables

Table S1. Names and sequences of oligonucleotides used in this work. L-Oligonucleotides are indicated in blue. /Cy5/ = cyanine5 dye. /FAM/ = 6-fluorescein. Asterisk indicates oligonucleotides that were prepared and characterized previously or purchased through IDT.[2]

Sequence Name	Sequence Identity 5′→3′	DNA/RNA
*A_{30}	AAAAAAAAAAAAAAAAAAAAAAAAAAAAAA -/Cy5/	RNA
C_{30}	CCCCCCCCCCCCCCCCCCCCCCCCCCCCCC -/Cy5/	RNA
U_{30}	UUUUUUUUUUUUUUUUUUUUUUUUUUUUUU -/Cy5/	DNA
G_{30}	GGGGGGGGGGGGGGGGGGGGGGGGGGGGGG -/Cy5/	RNA
*$(GGAA)_{10}$	GGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAA -/Cy5/	RNA
*L-$(GGAA)_{10}$	GGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAA -/Cy5/	RNA
*$(G_3A_4)_4$	GGGAAAAGGGAAAAGGGAAAAGGGAAAA -/Cy5/	RNA
$(GA)_{20}$	GAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGA-/Cy5/	RNA
$(GA)_{15}$	GAGAGAGAGAGAGAGAGAGAGAGAGAGAGA-/Cy5/	RNA
$(GA)_{10}$	GAGAGAGAGAGAGAGAGAGA-/Cy5/	RNA
$(GA)_5$	GAGAGAGAGA-/Cy5/	RNA
*GC_{HP}	GCGCGCGCGCGCGCGAGAGCGCGCGCGCGCGC-/Cy5/	RNA
MUT_{HP}	GCGCGCGCGCGCGCGAGAGCGCGAAAACGCGGGAGG--/Cy5/	RNA
$(GU)_{15}$	GUGUGUGUGUGUGUGUGUGUGUGUGU-/Cy5/	RNA
*$(GA)_{20}$	GAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGA-/Cy5/	DNA
BZ-dTFWD	/FAM/-TGAGGATGTATATATCTGA/dT/GCGCCGGTGGAGC	DNA
BZ-dUFWD	/FAM/-TGAGGATGTATATATCTGA/dU/GCGCCGGTGGAGC	DNA
BZ-FUFWD	/FAM/-TGAGGATGTATATATCTGA/^{F}U/GCGCCGGTGGAGC	DNA
BZ-REV	GCTCCACCGGCGCGTACGATATATACATCCTCA	DNA

Table S2. Names and sequences of primers used for HOTAIR and TFF1e DNAs and RNAs.

Sequence Name	Sequence Identity 5′→3′	DNA/RNA
TFF1eFWD	TTCTAATACGACTCACTATAGGGCTTTTGACCCTAAGGTCCCTT	DNA
TFF1eREV	CTCTCATCCCTGGCTCCCAG	DNA
HOTAIRFWD	TTCTAATACGACTCACTATAGAGCCCAGAGTTACAGACGG	DNA
HOTAIRREV	GCCCCTCCTTCCTCTCGCCGCCGTCTGTAACTCTGGGCTC	DNA

Table S3. Names and sequences of lncRNA fragments and the corresponding DNA templates used to prepared them. T7 promoter sequence is underlined.

Name	Sequence
TFF1e-DNA	TTCTAATACGACTCACTATAGGGCTTTTGACCCTAAGGTCCCTTAAATGCA ACCCTGGTCTTGCTATAAAAACCAGGGCGTGACCCAGTGGTTGTCAATTC AGCAGCAGTGACTTGGTGGCTCCAGGGTGTCAGACGCTTGTGCTGAAAA CAGAGATCACAGCTAGTCCTGGGCAGCTCTGGGTGGGTGCAGCCTGGCA GGCAGAGGAGCTGGGGCCCGGGAGGAAGAAGGAGCCTCACGACATGGG AAAGGAGGAGGCAGCGAGGAGGAGCCCCTGCTGGGATGGGGATGGGTC GGGCGTGCCCCGTGTGCCAGGAGCAGGGAGAAAGTGGGAGGGGGCGG GGGGCTGAGACAGTGCGGGAGAGGACCTTGCCTGCCTGCTGAGGGGCT GGAACCTCGCTGGGAGCCAGGGATGAGAG
TFF1e RNA	GGGCUUUUGACCCUAAGGUCCCUUAAAUGCAACCCUGGUCUUGCUAUA AAAACCAGGGCGUGACCCAGUGGUUGUCAAUUCAGCAGCAGUGACUUG GUGGCUCCAGGGUGUCAGACGCUUGUGCUGAAAACAGAGAUCACAGCU AGUCCUGGGCAGCUCUGGGUGGGUGCAGCCUGGCAGGCAGAGGAGCU GGGGCCCGGGAGGAAGAAGGAGCCUCACGACAUGGGAAAGGAGGAGG CAGCGAGGAGGAGCCCCUGCUGGGAUGGGGAUGGGUCGGGCGUGCCC CGUGUGCCAGGAGCAGGGAGAAAGUGGGAGGGGGCGGGGGGCUGAGA CAGUGCGGGAGAGGACCUUGCCUGCCUGCUGAGGGGCUGGAACCUCG CUGGGAGCCAGGGAUGAGAG
HOTAIR-DNA	TTCTAATACGACTCACTATAGGACTCGCCTGTGCTCTGGAGCTTGATCCG AAAGCTTCCACAGTGAGGACTGCTCCGTGGGGGTAAGAGAGCACCAGGC ACTGAGGCCTGGGAGTTCCACAGACCAACACCCCTGCTCCTGGCGGCTC

	CCACCCGGGGCTTAGACCCTCAGGTCCCTAATATCCCGGAGGTGCTCTC AATCAGAAAGGTCCTGCTCCGTTCGCAGTGGAATGGAACGGATTTAGAAG CCTGCAGTAGGGGAGTGGGGAGTGGAGAGAGGGAGCCCAGAGTTACAG ACGGCGGCGAGAGGAAGGAGGGGC
HOTAIR RNA	GGACUCGCCUGUGCUCUGGAGCUUGAUCCGAAAGCUUCCACAGUGAG GACUGCUCCGUGGGGGUAAGAGAGCACCAGGCACUGAGGCCUGGGAG UUCCACAGACCAACACCCCUGCUCCUGGCGGCUCCCACCCGGGGCUUA GACCCUCAGGUCCCUAAUAUCCCGGAGGUGCUCUCAAUCAGAAAGGUC CUGCUCCGUUCGCAGUGGAAUGGAACGGAUUUAGAAGCCUGCAGUAGG GGAGUGGGGAGUGGAGAGAGGGAGCCCAGAGUUACAGACGGCGGCGA GAGGAAGGAGGGGC

S3. References.

1. Maiti A., Noon M.S., MacKerell A.D. Jr., Pozharski E., Drohat A.C. Lesion processing by a repair enzyme is severely curtailed by residues needed to prevent aberrant activity on undamaged DNA. *Proc. Natl. Acad. Sci. U S A.* 2012; 109, 8091-6.
2. Deckard C.E., Sczepanski J.T. Polycomb repressive complex 2 binds RNA irrespective of stereochemistry. *Chem. Commun.* 2018; 54, 12061-12064.