Supplementary Material

3’ end additions by T7 RNA polymerase are RNA self-templated, distributive, and diverse in character – RNA-Seq analyses

Yasaman Gholamalipour[1], Aruni Karunanayake Mudiyanselage[1], and Craig T. Martin[1,*]

[1] Department of Chemistry, University of Massachusetts Amherst, Amherst, MA, 01003, USA
* To whom correspondence should be addressed. Tel: +1 413 545 3299; Fax: +1 413 545 4490;

Email: cmartin@chem.umass.edu

Table of contents:

Table S1. Sequences used for transcription in this study

DNA	Sequences
Non-template	5'-AATTAATACGACTCACTATAGG-3'
5N template	3'-TTAATTATGCTGAGTGATATCCNNNNNCATCTCCACTTCTAAAT-5'
5N-U→A template	3'-TTAATTATGCTGAGTGATATCCNNNNNCATCTCCTCTTCTAAAT-5'
5N-UG→AC template	3'-TTAATTATGCTGAGTGATATCCNNNNNCATCTCCTGTTCTAAAT-5'
3N template	3'-TTAATTATGCTGAGTGATATCCTCNNNGATGCAGCTGCGTAAAT-5'
3N-G→U template	3'-TTAATTATGCTGAGTGATATCCTCNNNGATGCAGATGCGTAAAT-5'
Synthetic 24mer RNA	5'-GGAAUAAGUAGAGGUGAAGAUUUA-3'

T7 promoter sequences are underlined

Table S2. Sequences for RNA-Seq library preparation

Oligo for:	Name	Sequences (5' to 3' direction)
3'Adapter Ligation	3' adapter (all)	P-TGGAATTCTCGGGTGCCAAGG-Biotin
5' Adapter Ligation	5' adapter for 3N (high yield)	GUUCAGAGUUCUACAGUCCGACGAUCUAAUCA
	5' adapter for 5N (low yield)	GUUCAGAGUUCUACAGUCCGACGAUCUACUAU
	5' adapter for 5N (high yield)	GUUCAGAGUUCUACAGUCCGACGAUCACAUAA
	5' adapter for synthetic RNA (low yield)	GUUCAGAGUUCUACAGUCCGACGAUCAUAUCC
Reverse Transcription	RT-Primer (all)	GCCTTGGCACCCGAGAATTCCA

Designed barcode sequences for each 5' adapter are underlined

Table S3. Illumina primers for TruSeq Small RNA Library

Oligo for:	Application	Sequences (5' to 3' direction)
Forward Primer	all	AATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTACAGTCCGA
Reverse Primer	3N (high yield)	CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA
	5N (low yield)	CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA
	5N (high yield)	CAAGCAGAAGACGGCATACGAGATGGAACTGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA
	Synthetic RNA (low yield)	CAAGCAGAAGACGGCATACGAGATCTCTACGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA

Table S4. Summary of sequence data populations for each RNA-Seq data set.

Experiments were multiplexed (typically, 10 per sequencing run), as a single sequencing run yields far more sequence reads than necessary for this study. The first data column below is the actual percentage that the indicated sequence set represents in the larger sequencing run. Runs were not multiplexed equally, as sequences with more complexity (e.g. randomization) warrant deeper reads.

The remaining columns are labeled as in the lower right of the Figure S1 flow chart, and as in that chart, follow on each other, left to right. "Raw Seq Reads" were trimmed of adapters (only sequences with *both* 5' and 3' adapters were retained), then filtered to remove primer dimers and single base inserts, and were finally filtered to analyze only reads that begin with the expected initial sequence GG.

For the promoter driven transcription reactions (*i.e.*, all except "24mer"), abortive products represent the majority of captured products ("GG starts"), as typically observed in gel electrophoresis. As they are not the focus of this study, they were removed by filtering for only sequences 15 bases in length or longer. For 3' end analyses, sequences were further aligned to an encoded upstream sequence beginning at position +10, which adjusts for mis-initiation or slippage in the first few bases of the transcript.

	Percent in Multiplex	Raw Seq Reads	Adapter Trimmed	Trimmed	GG starts	≥15mer	≥15mer & align
5N (low yield)	6.9%	36,026	24,752	23,706	20,626	6,193	5,905
5N (high yield)	25.2%	130,900	91,932	85,488	47,356	19,283	17,350
5N (high yield) replicate	9.9%	92,929	74,784	70,973	33,967	14,725	13,412
24mer (low yield)	3.9%	37,206	25,322	21,880	1,442	1,218	1,162
3N (high yield)	10.3%	96,796	71,068	67,112	33,353	7,702	7,177

Mapped reads '≥15mer' have been deposited in the Small Read Archive (SRA) (http://www.ncbi.nlm.nih.gov/sra) with the BioProject accession code PRJNA486161, with entries SAMN09839052 (5N low yield), SAMN09839053 (5N high yield), SAMN09839054 (5N high yield, Replicate), SAMN09839055 (24mer low yield), SAMN09839056 (3N high yield).

Figure S1. Flow chart of *in vitro* transcription, followed by RNA-Seq data analysis.

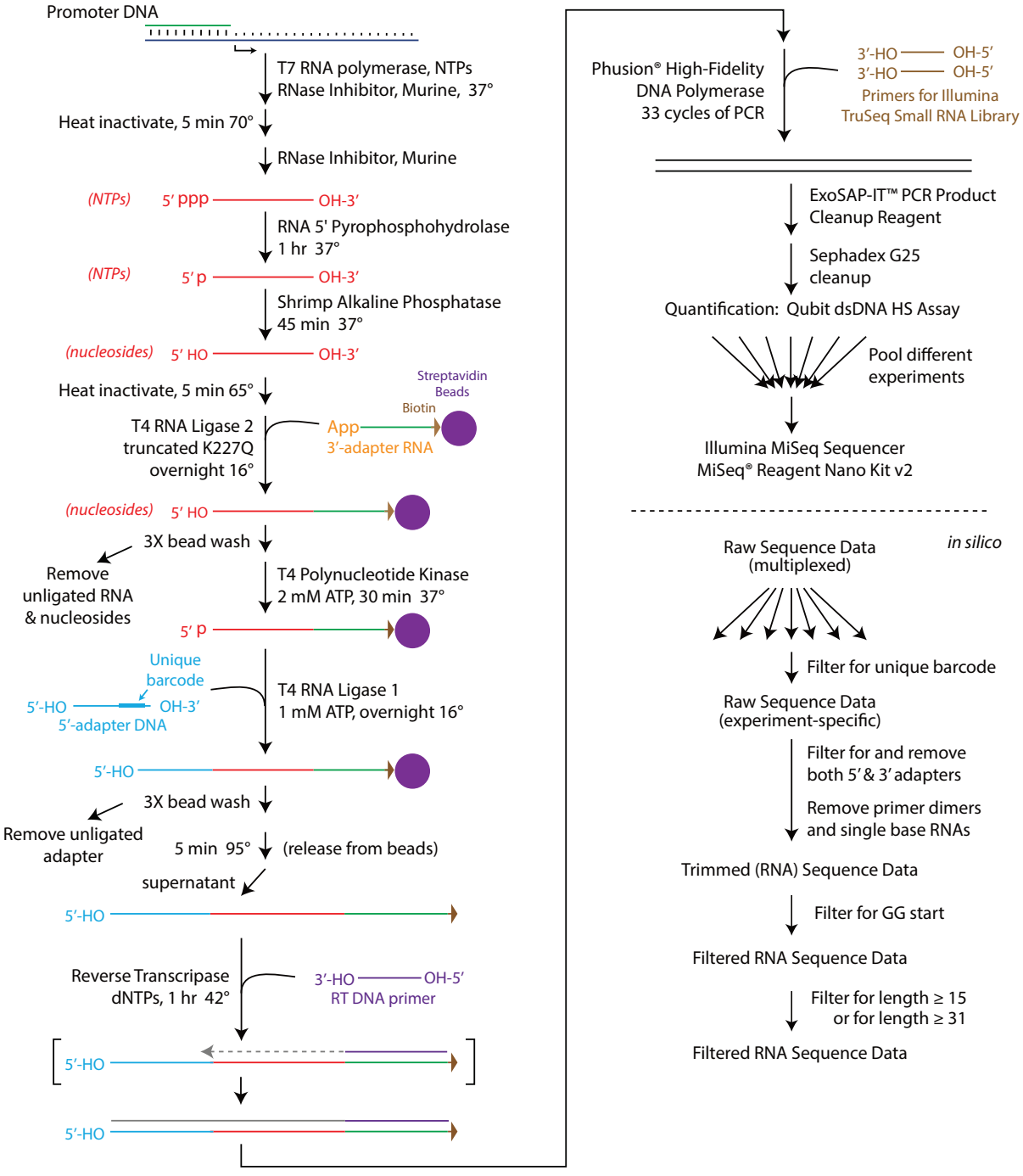

Figure S2. Comparison of the most abundant 3' end heterogeneities. A) the same data as in Figure 1C, high yield, but extended to sequences representing ≥0.5% of the pool and showing the actual counts of each sequence; B) a replicate run of that experiment, indicating very good reproducibility; C) the same data as in Figure 1C, low yield; D) the same data as in Figure 5D, using template 3N, which encodes a different sequence upstream of position +20. Comparison of (A or C) and (D) demonstrates that the identities of the 3' additions change significantly. The total number of sequence reads in the pool are shown at the bottom of each set.

The reader is cautioned that while the generally good agreement between replicates A and B indicates a high level of reproducibility in the assay, factors such as (reproducible) ligation bias likely contribute more to uncertainties and should limit subtle interpretations of the data.

A

5N high yield conditions

		Encoded: AGAGGUGAAGAUUUA
1520	8.8%	AGAGGUGAAGAUUUAC
1293	7.5%	AGAGGUGAAGAUUUA
1214	7.0%	AGAGGUGAAGAUUUACC
558	3.2%	AGAGGUGAAGAUUUACA
522	3.0%	AGAGGUGAAGAUUUACU
361	2.1%	AGAGGUGAAGAUU
335	1.9%	AGAGGUGAAGAUUUACCU
335	1.9%	AGAGGUGAAGAUUUAAC
320	1.8%	AGAGGUGAAGAUUU
312	1.8%	AGAGGUGAAGAUUUACACC
253	1.5%	AGAGGUGAAGAUUUC
247	1.4%	AGAGGUGAAGAUUUACUU
242	1.4%	AGAGGUGAAGAUUUACCUCUACU
239	1.4%	AGAGGU
222	1.3%	AGAGGUGAAGAUUUACCUCUAC
220	1.3%	AGAGGUGAAGAUUUACAC
202	1.2%	AGAGGUGAAGAUUUACCUCUA
200	1.2%	AGAGGUGAAGAUUUACCUCUACA
199	1.1%	AGAGGUGAAGAUUUACCUC
181	1.0%	AGAGGUGAAGAUUUACUC
180	1.0%	AGAGGUGAAGA
178	1.0%	AGAGGUG
151	0.9%	AGAGGUGAAGAUUUACACCUCUA
144	0.8%	AGAGGUGAAGAUUUACUUC
136	0.8%	AGAGGUGAAGAUUUCACC
136	0.8%	AGAGGUGAAGAU
134	0.8%	AGAGGUGAAGAUUUACUA
127	0.7%	AGAGGUGAAGAUUUACACCU
108	0.6%	AGAGGUGAAGAUUUACACCUCUAC
105	0.6%	AGAGGUGAAGAUUUAA
100	0.6%	AGAGGUGAAGAUUUACG
90	0.5%	AGAGGUGAAGAUUUCAC

17350 Sequences ≥15 bases long and aligning to the sequence 'AGAGG'

B

5N high yield conditions - replicate

		Encoded: AGAGGUGAAGAUUUA
1346	10.0%	AGAGGUGAAGAUUUA
1002	7.5%	AGAGGUGAAGAUUUAC
941	7.0%	AGAGGUGAAGAUUUACC
529	3.9%	AGAGGUGAAGAUUUACA
395	2.9%	AGAGGUGAAGAUUUACU
342	2.5%	AGAGGUGAAGAUUU
315	2.3%	AGAGGUGAAGAUUUACCU
242	1.8%	AGAGGUGAAGAUU
233	1.7%	AGAGGUGAAGAUUUACUU
208	1.6%	AGAGGUGAAGAUUUACACC
198	1.5%	AGAGGUGAAGAUUUC
185	1.4%	AGAGGUGAAGAUUUACCUC
174	1.3%	AGAGGUGAAGAUUUAAC
167	1.2%	AGAGGUG
164	1.2%	AGAGGUGAAGAUUUACCUCUACU
154	1.1%	AGAGGUGAAGA
154	1.1%	AGAGGUGAAGAUUUACCUCUA
146	1.1%	AGAGGUGAAGAUUUACAC
143	1.1%	AGAGGU
143	1.1%	AGAGGUGAAGAUUUACUC
140	1.0%	AGAGGUGAAGAUUUACCUCUACA
127	0.9%	AGAGGUGAAGAUUUACUUC
111	0.8%	AGAGGUGAAGAUUUACUA
111	0.8%	AGAGGUGAAGAUUUACACCUCUA
109	0.8%	AGAGGUGAAGAUUUACG
108	0.8%	AGAGGUGAAGAUUUACCUCUAC
104	0.8%	AGAGGUGAAGAUUUACACCU
100	0.7%	AGAGGUGAAGAUUUCACC
95	0.7%	AGAGGUGAAGAUUUAA
85	0.6%	AGAGGUGAAGAUUUAU
83	0.6%	AGAGGUGAAG
81	0.6%	AGAGGUGAAGAU
75	0.6%	AGAGGUGAAGAUUUACCC
72	0.5%	AGAGGUGAAGAUUUAUU

13412 Sequences ≥15 bases long and aligning to the sequence 'AGAGG'

C

5N low yield conditions

		Encoded: AGAGGUGAAGAUUUA
1744	29.5%	AGAGGUGAAGAUUUA
1108	18.8%	AGAGGUGAAGAUUUAC
728	12.3%	AGAGGUGAAGAUUU
231	3.9%	AGAGGUGAAGAUU
187	3.2%	AGAGGUGAAGA
185	3.1%	AGAGGUGAAGAUUUACC
169	2.9%	AGAGGUGAAGAU
141	2.4%	AGAGGUG
127	2.2%	AGAGGUGAAGAUUUAA
102	1.7%	AGAGGUGAAGAUUUAU
92	1.6%	AGAGGUGAAG
79	1.3%	AGAGGUGAAGAUUUACA
75	1.3%	AGAGGU
72	1.2%	AGAGGUGAA
69	1.2%	AGAGGUGAAGAUUUACU
61	1.0%	AGAGGUGAAGAUUUAAC
44	0.7%	AGAGGUGAAGAUUUC
36	0.6%	AGAGGUGA
32	0.5%	AGAGGUGAAGAUUUACG

5905 Sequences ≥15 bases long and aligning to the sequence 'AGAGG'

D

3N high yield conditions

		Encoded: ACGUCGACGCAUUUA
988	13.8%	ACGUCGACGCAUUUAC
837	11.7%	ACGUCGACGCAUUUAG
639	8.9%	ACGUCGACGCAUUUA
456	6.4%	ACGUCGACGCAUUU
432	6.0%	ACGUCGACGC
411	5.7%	ACGUCGACGCAUU
281	3.9%	ACGUCGACGCA
242	3.4%	ACGUCGA
239	3.3%	ACGUCGAC
185	2.6%	ACGUCGACGCAUUUAA
180	2.5%	ACGUCGACGCAUUUAGA
175	2.4%	ACGUCGACG
166	2.3%	ACGUCGACGCAUUUAGC
109	1.5%	ACGUCGACGCAUUUACC
98	1.4%	ACGUCGACGCAUUUAU
92	1.3%	ACGUCG
59	0.8%	ACGUCGACGCAU
38	0.5%	ACGUCGACGCAUUUACU

7177 Sequences ≥15 bases long and aligning to the sequence 'ACGUC'

Figure S3. Estimation of *cis* vs *trans* initiated distributions. In order to assess whether a transcript initiated from self-templating (*cis*) vs templating from another RNA (*trans*), we utilized the fact that the original DNA template encodes randomized bases from position +3 through +7. Following the flow chart below, data were filtered for only RNAs that read into that key sequence region, plus at least two bases beyond. A sequence is tagged as *cis* only if the (entire) sequence past the key sequence is the exact inverse complement of the corresponding region of the initial sequence.
Note that the requirement for an exact sequence match likely *underestimates* the percentage of reactions that were *cis* in origin. True *cis*-derived transcripts that subsequently (distributively) add even one additional base are tagged as *trans*.

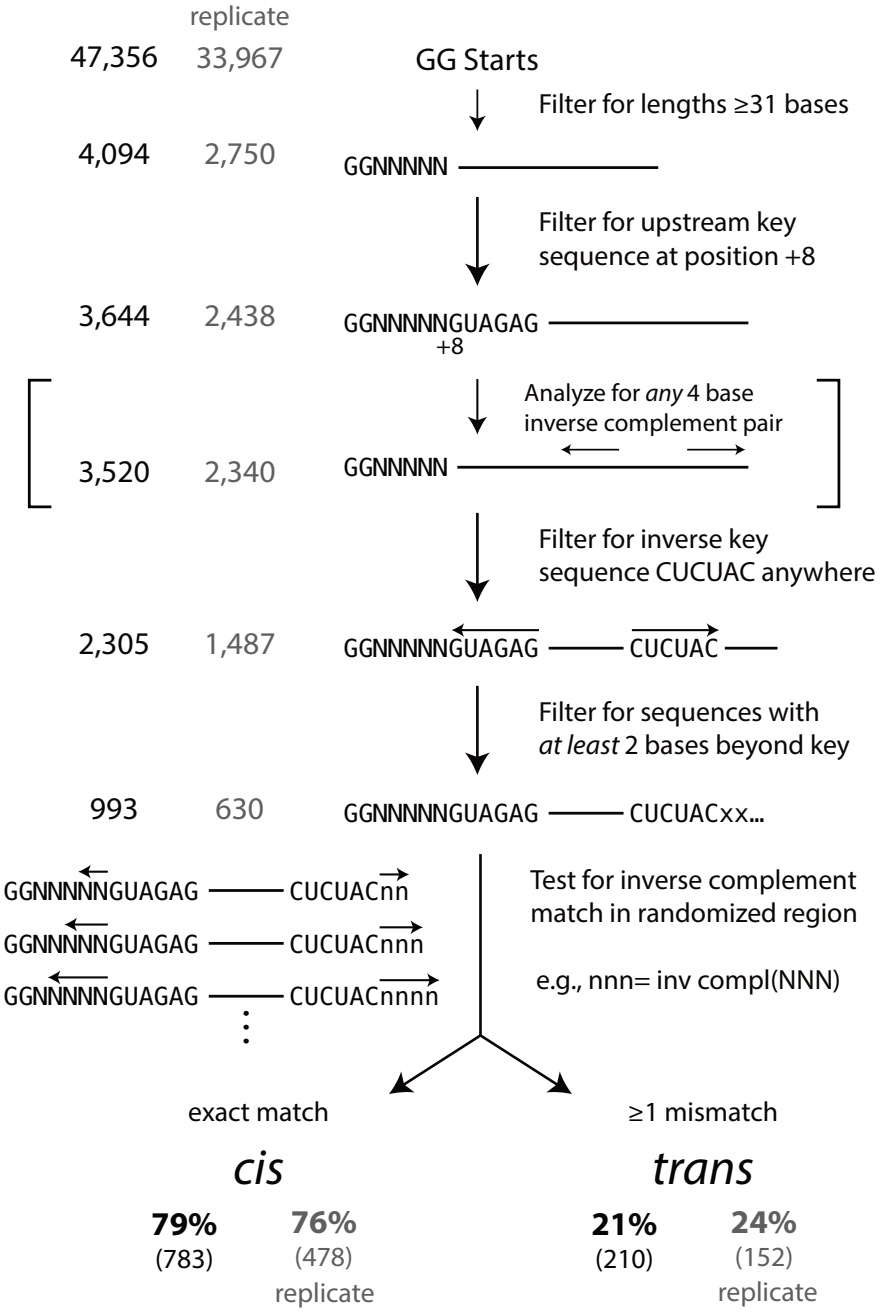

Figure S4. Lane profile traces of electrophoretic data. In order to allow more quantitative comparisons of the RNA gels in figures 5B and 6B, the following analyzes each lane for the relative amounts of each product (or range of products). As lanes were loaded differently, each trace is normalized to itself, allowing assessment of the relative amounts of encoded 24mer and longer primer-extended RNAs. Gels were analyzed using ImageJ v1.51. Note that "High 5N" and "High 3N"in Figure 5B and "5N" and "3N" in Figure 6B serve effectively as replicates, respectively.