Supplementary information

Characterization and implementation of the MarathonRT template-switching reaction to expand the capabilities of RNA-Seq

Li-Tao Guo[a], Anastasiya Grinko[b], Sara Olson[c], Alexander M. Leipold[b,d], Brenton Graveley[c], Antoine-Emmanuel Saliba[b,d], and Anna Marie Pyle[a,e,f,*]

[a] *Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven CT 06520, USA*

[b] *Helmholtz Institute for RNA-based Infection Research (HIRI), Helmholtz-Centre for Infection Research (HZI), Würzburg, Germany*

[c] *Genetics and Genome Sciences, University of Connecticut Health, Farmington CT 06030, USA*

[d] *University of Würzburg, Faculty of Medicine, Institute of Molecular Infection Biology (IMIB), Würzburg, Germany*

[e] *Department of Chemistry, Yale University, New Haven CT 06520, USA*

[f] *Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA*

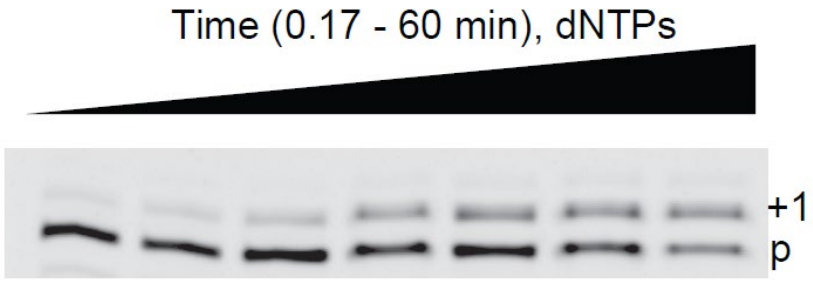
*Correspondence should be addressed to Anna Marie Pyle: anna.pyle@yale.edu

Supplementary Table 1: RNA and DNA oligonucleotides used in this work.

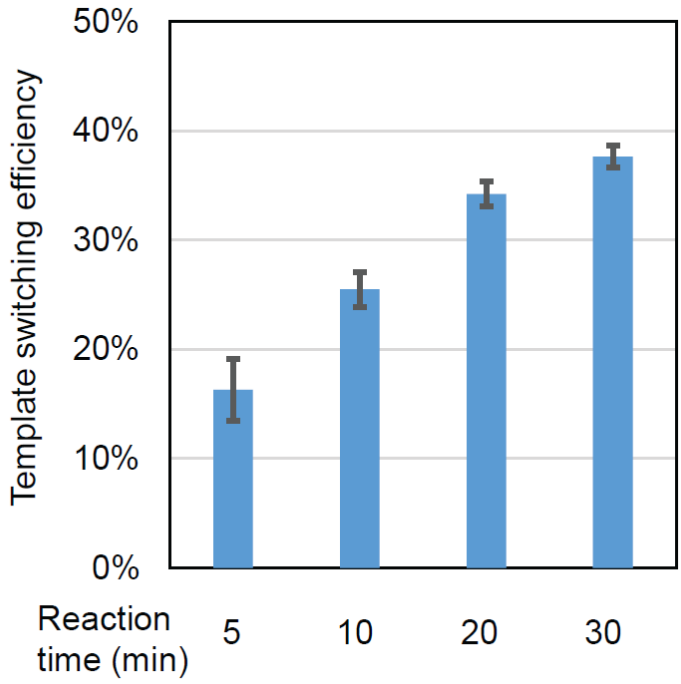
RNA oligo	Sequence	Length	Note
GCrA30	5'-GCAAAAA AAAAA AAAAA AAAAA AAAAA AAAAA-SpC3-3'	32 nt	Containing 5'-OH and 3' C3 Spacer
AGCrA25	5'-ppp-AGCAAAAA AAAAA AAAAA AAAAA AAAAA-ddC-3'	29 nt	Containing 5'-triphophate and 2',3'-Dideoxycytidine
CGCrA25	5'-ppp-CGCAAAAA AAAAA AAAAA AAAAA AAAAA-ddC-3'	29 nt	Containing 5'-triphophate and 2',3'-Dideoxycytidine
GGCrA25	5'-ppp-GGCAAAAA AAAAA AAAAA AAAAA AAAAA-ddC-3'	29 nt	Containing 5'-triphophate and 2',3'-Dideoxycytidine
UGCrA25	5'-ppp-UGCAAAAA AAAAA AAAAA AAAAA AAAAA-ddC-3'	29 nt	Containing 5'-triphophate and 2',3'-Dideoxycytidine
RNA TSOs			
MRT_TSO_CAA	5'-AGAGACAGAUUGCGCAAUGACAA-3'	23 nt	
MRT_TSO_CAG	5'-AGAGACAGAUUGCGCAAUGACAG-3'	23 nt	
MRT_TSO_CAU	5'-AGAGACAGAUUGCGCAAUGACAU-3'	23 nt	
MRT_TSO_CAC	5'-AGAGACAGAUUGCGCAAUGACAC-3'	23 nt	
MRT_TSO_UUU	5'-AGAGACAGAUUGCGCAAUGACAUUU-3'	25 nt	
DNA Primer			
FAM_dT30GC	5'-FAM-TTTTT TTTTT TTTTT TTTTT TTTTT TTTTTGC-3'	32 nt	FAM labeling at 5'-end
FAM_dT30GCG	5'-FAM-TTTTT TTTTT TTTTT TTTTT TTTTT TTTTTGCG-3'	33 nt	FAM labeling at 5'-end
FAM_dT30GCAAA	5'-FAM-TTTTT TTTTT TTTTT TTTTT TTTTT TTTTTGCAAA-3'	35 nt	FAM labeling at 5'-end
FAM_dT25GC	5'-FAM-TTTTT TTTTT TTTTT TTTTT TTTTTGC-3'	27 nt	FAM labeling at 5'-end
MRT_dT18_opt	5'-CCTTCTCCTTCTCCTCCTTTCTCCTTTTTTTTTTTTTTTTTT-3'	42 nt	RT primer with optimized sequence for library construction
AmpPCR	5'-CCCTCTCTCTCTCTTTCCTCTCTC-3'	24 nt	Primer used to amplify full-length uMRT cDNA library
MLV_dT30	5'-AAGCAGTGGTATCAACGCAGAGTGAAT TTTTT TTTTT TTTTT TTTTT TTTTT TTTTT-3'	57 nt	The RT primer for SSII and Maxima
mlvPCR	5′-AAGCAGTGGTATCAACGCAGAGT-3′	23 nt	Primer used to amplify SSII and Maxima cDNA libraries
DNA TSOs			
MRT_TSO_opt	5'-CCCTCTCTCTCTCTTTCCTCTCTCTTTT-3'	28 nt	Optimized sequence for library construction
MRT_TSO_5ab	5'-5abasic-CCCTCTCTCTCTCTTTCCTCTCTCTTTT-3'	28 nt	TSO with 5 abasic sites at the 5′-end
MRT_TSO_3ab	5'-3abasic-CCCTCTCTCTCTCTTTCCTCTCTCTTTT-3'	28 nt	TSO with 3 abasic sites at the 5′-end
MRT_TSO_trebler	5'-trebler-CCCTCTCTCTCTCTTTCCTCTCTCTTTT-3'	28 nt	TSO with 5′-trebler modification
MRT_TSO_trityl	5'-trityl-CCCTCTCTCTCTCTTTCCTCTCTCTTTT-3'	28 nt	TSO with 5′-trityl modification
MLV_TSO	5'-AAGCAGTGGTATCAACGCAGAGTGAATrGrG+G-3'	32 nt	The TSO for SSII and Maxima

Supplementary Table 2. Summary of the yield of total reads and overall alignment rate to SIRV-Set 3 templates in the sequencing experiment using different TSOs, including the TSO with no 5'-modification and TSOs containing five 5'-abasic sites, three 5'-abasic sites, 5'-trebler and 5'-trityl.

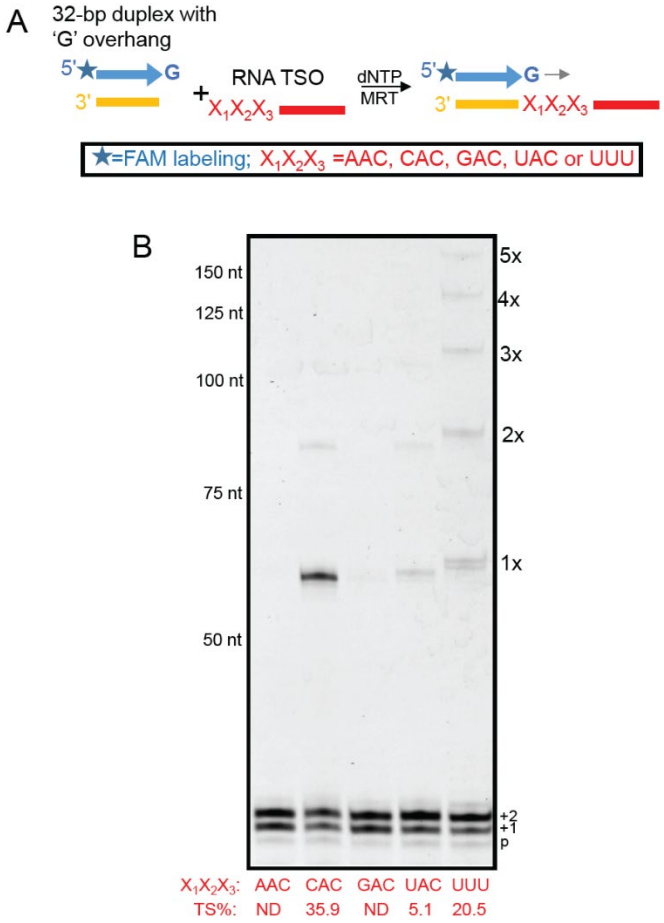
TSO	**Total reads (paired end)**	**Overall alignment rate**
no 5'-mod	587,745	2.35%
5'-5absic	219,643	95.2%
5'-3absic	640,085	95.3%
5'-trebler	513,231	95.8%
5'-trityl	597,837	94.8%



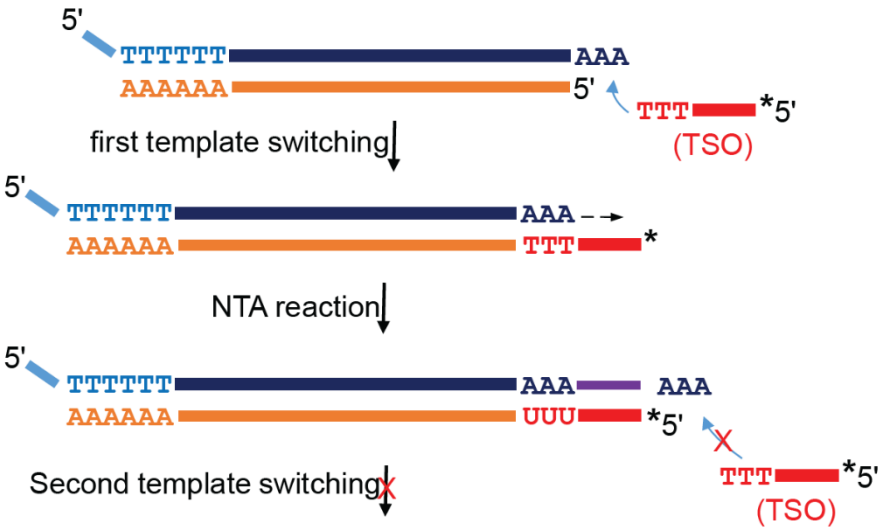
Suppl. Fig. 1. A representative gel figure showing the NTA products by TGIRT in time course. An equimolar dNTP mix was used in this experiment. The reaction time spans from 10 sec to 60 min.



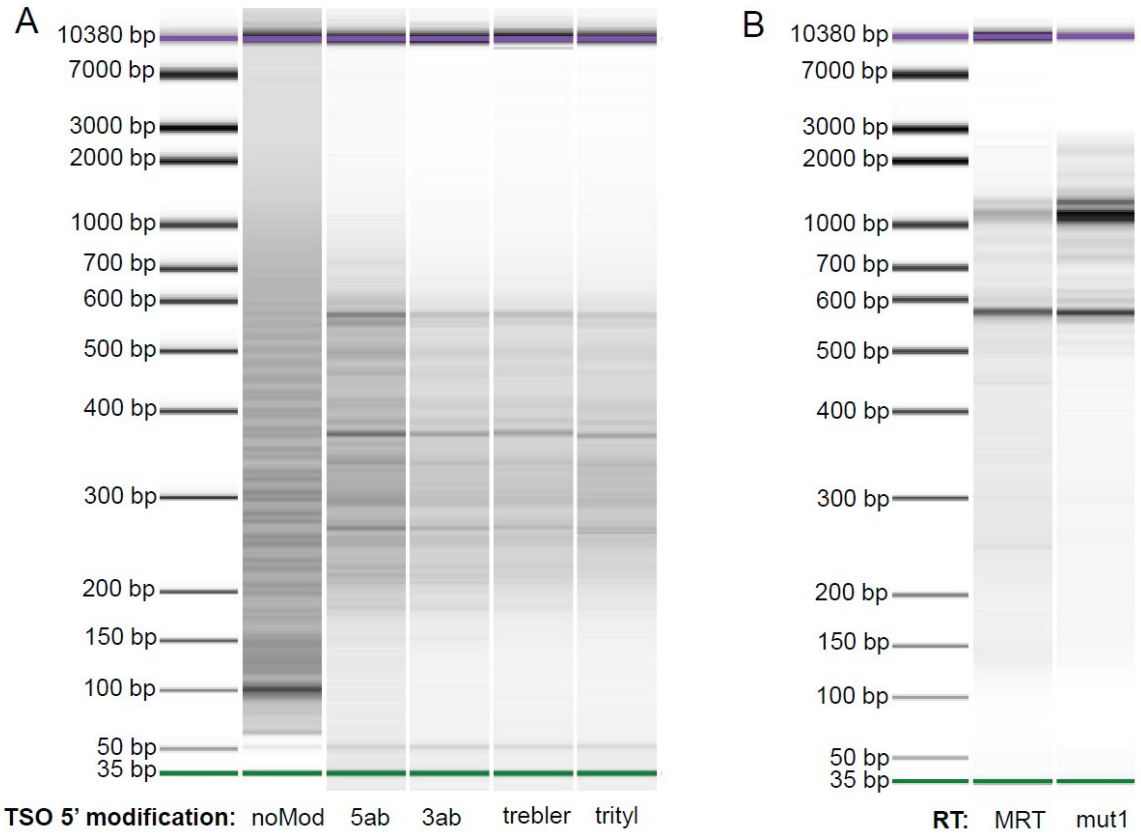
Suppl. Figure 2. Time-course experiment to monitor the template switching efficiency of MarathonRT. Each reaction was repeated three times with mean and standard deviation shown in the figures.



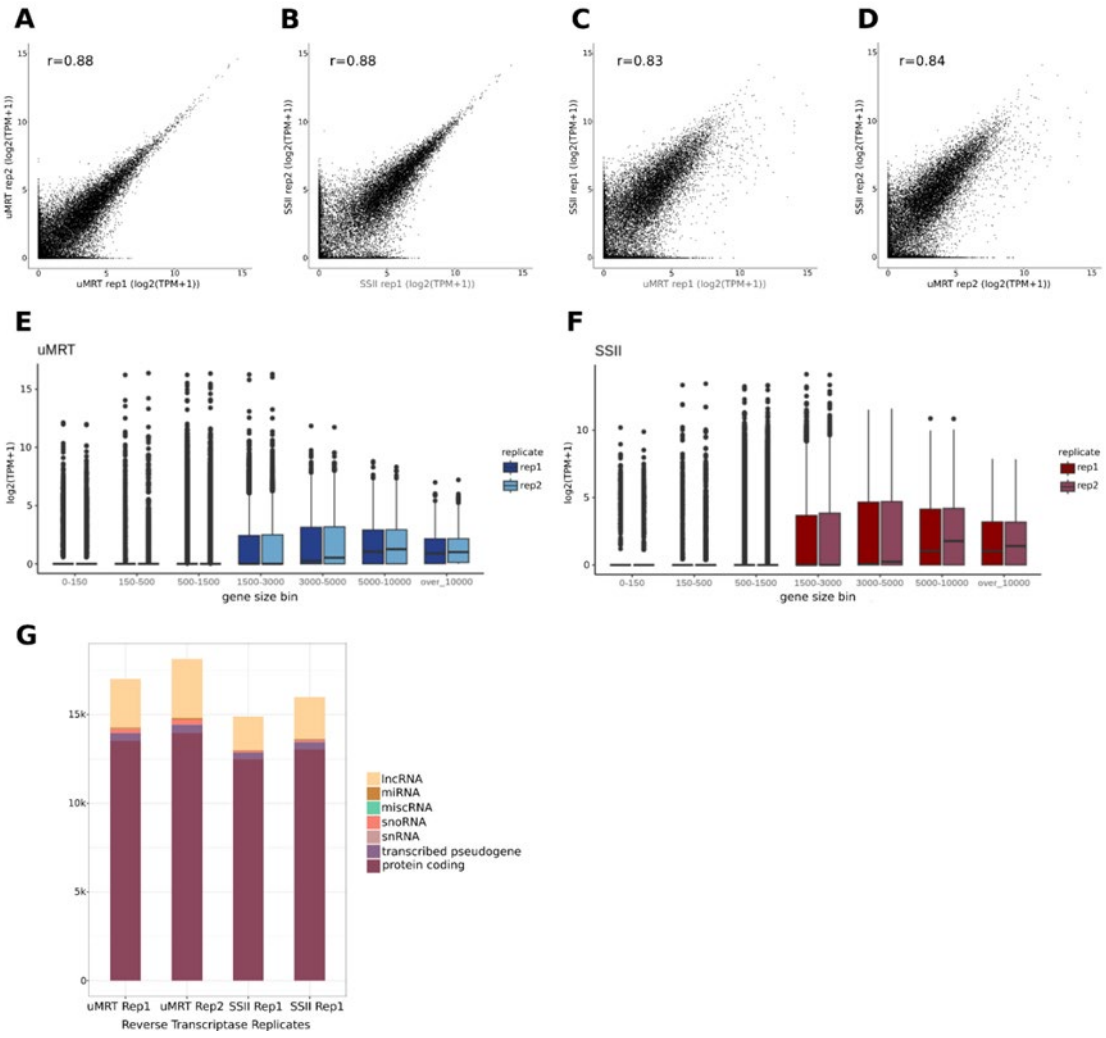
Suppl. Figure 3. Examining the TSO annealing specificity using the DNA/RNA duplex with a single G overhang in combination with the set of TSOs as shown in **Fig. 4**. (A) Outline of the reaction setup. (B) A representative gel figure showing the template switching products generated in the reactions. The template switching efficiency (TS%) for each reaction was calculated from three independent experiments shown in the figure. ND, not detected.



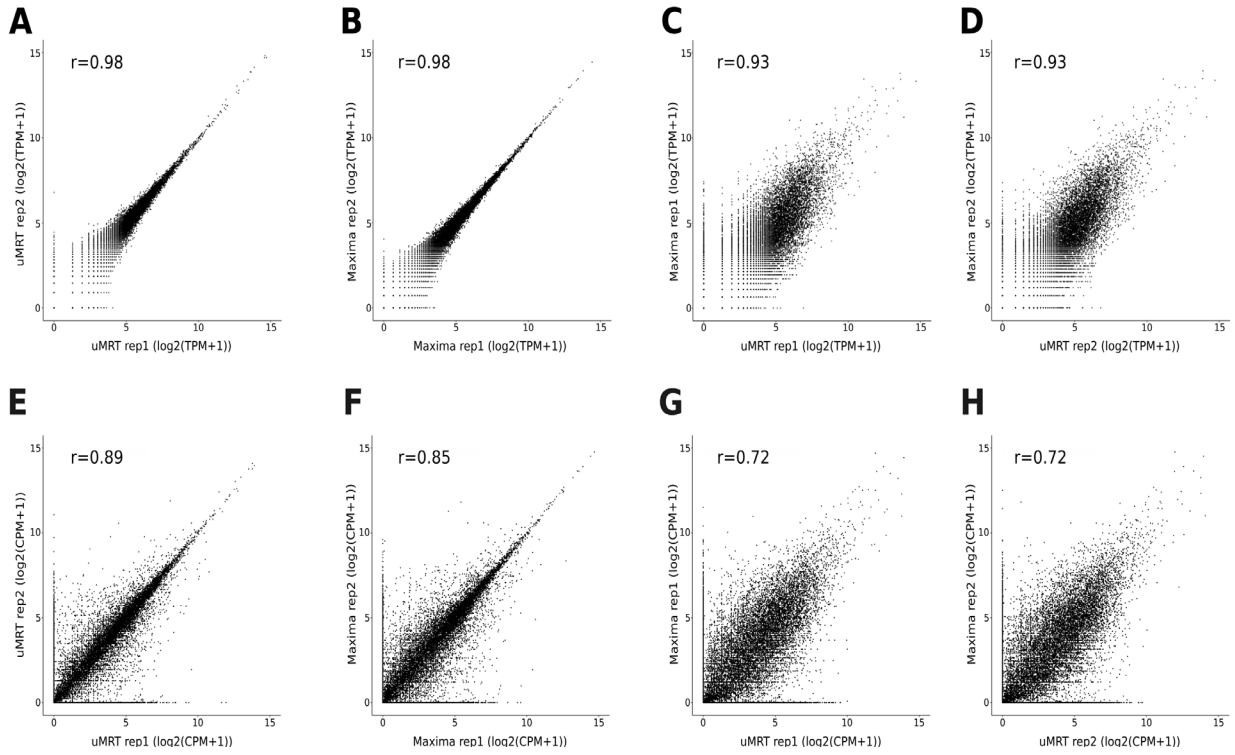
Suppl. Figure 4. Schematic outline showing the prevention of TSO concatenation by introducing a bulky chemical group (*) to the 5'-OH of TSO. Four types of chemical modifications were tested, including five or three consecutive abasic sites, a trebler and a trityl.



Suppl. Figure 5. Length distribution of pre-amplified cDNA libraries from SIRV-Set3 RNA transcripts generated by template switching reaction of MarathonRT or its variant mut1. (A) BioAnalyzer gel image showing the profiles of pre-amplified SIRV-Set3 cDNA libraries generated using MRT and the TSOs without 5' modification (noMod), with 5 consecutive abasic sites (5ab), 3 consecutive abasic sites (3ab), a trebler and a trityl chemical group respectively at the 5' end. The libraries were generated using the one-step reaction (see Methods), in which 0.1 ng SIRV-Set3 transcripts, 0.25 µM MRT_dT18_opt primer, unbalanced dNTP mix (0.8 mM dATP and 0.4 mM of the other three dNTPs), 250 nM MarathonRT, 1 µM TSO, and the optimized template switching buffer (50 mM Tris-HCl pH 8.3, 100 mM KCl, 2 mM $MgCl_2$, 5 mM DTT, 20% glycerol and 12% PEG4000) were mixed together to prepare a 10-µL reaction for each TSO. The reactions were carried out by incubation at 42°C for 90 min, followed by heat inactivation and library preamplification by PCR. (B) BioAnalyzer image of the pre-amplified SIRV-Set3 cDNA libraries generated by MRT and mut1 using two-step reverse transcription/template switching reaction. The TSO used in this experiment has 5 consecutive abasic sites (5ab) at its 5' end.



Suppl. Figure 6. Assessment of reproducibility of short-read RNA-seq libraries prepared using template switching reactions of uMRT and SSII using human universal reference total RNA as template. (A,B) Scatterplot comparison of TPM normalized gene counts (log2(TPM+1) scale) between replicate samples prepared with uMRT (A) and SSII (B), respectively. (C,D) Scatterplot comparison of TPM normalized gene counts (log2(TPM+1) scale) between samples prepared with uMRT and SSII for replicate 1 (C) and replicate 2 (D), respectively. (E,F) Boxplots of log2(TPM+1) of genes divided into gene size bins across replicate 1 and replicate 2 prepared with uMRT (E) and SSII (F), respectively. Center line indicates the median, box limits indicate the upper and lower quantiles, and whiskers indicate the 1.5x interquartile range. Gene length is calculated based on the output of FeatureCounts (v2.0.6) and corresponds to the bases covered by exons based on the provided gene annotation. (G) Bar charts illustrating the number of genes detected (TPM>0) between every replicate of both uMRT and SSII, respectively, with reads downsampled to 20 million for every individual sample to account for sequencing depth. Colors indicate the gene biotype of the detected genes depicting the distribution across all detected biotypes.



Suppl. Figure 7. Assessment of reproducibility of ONT long-read RNA-seq libraries prepared using template switching reactions of uMRT and Maxima using poly(A)+ mRNA from Huh7.5 cells as template. Pearson correlation coefficient (r) was used to evaluate reproducibility. (A,B) Scatterplot comparison of TPM normalized gene counts (log2(TPM+1) scale) between replicates prepared with uMRT (A) and Maxima (B), respectively. (C, D) Scatterplot comparison of TPM normalized gene counts (log2(TPM+1) scale) between samples prepared with uMRT and Maxima for replicate 1 (C) and replicate 2 (D), respectively. (E,F) Scatterplot comparison of CPM normalized isoform counts (log2(CPM+1) scale) between replicates prepared with uMRT (E) and Maxima (F), respectively. (G.H) Scatterplot comparison of CPM normalized isoform counts (log2(CPM+1) scale) between samples prepared with uMRT and Maxima for replicate 1 (G) and replicate2 (H), respectively.

Supplementary Python scripts – Associated to Figure 6

File name: sam2count.py

The Python script was used to calculate the read counts aligned to each ERCC transcript reference sequence. The sequencing reads were generated by Illumina NextSeq sequencer using the sequencing libraries built from Lexogen SIRV-Set 3 synthetic RNA transcripts by MarathonRT mut1. The sequencing read count data were used to generate Figure 6.

Requirement: Python 3

The Python scripts requires SAM files as the input. To obtain the SAM files, the fastq files were aligned to the ERCC templates using Bowtie2 (version 2.4.2) with options –local, --no-unal and --no-mixed activated.

Usage:

1. For the input SAM file, replace the strings in “ERCC_example.sam” with your own sample names with suffix included.
2. Replace the strings in "ERCC_example_output.csv" with your own names, which is the ERCC read count output file in csv format.
2. Copy the Python script below to create the file sam2count.py.
3. Place your SAM file and the ERCC_info.csv (provided in the supplementary materials) in the same directory.
3. Run the script in the form of:

```
        python sam2count.py
################################################################################


with open("ERCC_info.csv") as ERCCsummary:
    ERCCinfo = ERCCsummary.readlines()


erccName = []
erccConc = {}
erccLength = {}
erccGCcontent ={}
for ERCC in ERCCinfo:
    erccName.append(ERCC.split(",")[0])
    erccConc[ERCC.split(",")[0]]=ERCC.split(",")[2].strip()
    erccLength[ERCC.split(",")[0]]=ERCC.split(",")[3].strip()
    erccGCcontent[ERCC.split(",")[0]] = ERCC.split(",")[4].strip()
#print(len(erccName))
```

```
with open("ERCC_example.sam") as samFile:
    samFilelines = samFile.readlines()
# print(samFilelines[1])
transcriptsReadCount = {}
n=0
for samFileline in samFilelines:
    if samFileline[0] != "@":
        n += 1  # calcuate the total counts of sequencing reads aligned to the templates.
        # print(samFileline.split()[2])
        if samFileline.split()[2] in transcriptsReadCount.keys():
            transcriptsReadCount[samFileline.split()[2]] += 1
        else:
            transcriptsReadCount[samFileline.split()[2]] = 1


f = open("ERCC_example_output.csv", "w")
f.write("ERCC_ID, Read_count, Length(nt), Concentration(amoles/ul), GC_content"+"\n")
missERCC = []
for erccID in erccName:
    if erccID in transcriptsReadCount.keys():
        f.write(",".join([erccID,str(transcriptsReadCount[erccID]),erccLength[erccID],erccConc[erccID],
erccGCcontent[erccID] + "\n"]))
    else:
        f.write(",".join([erccID,"0",erccLength[erccID],erccConc[erccID], erccGCcontent[erccID] + "\n"]))
        missERCC.append(erccID)
f.write("Total reads: " + str(n) + "\nMissing transcripts: " + ' '.join(missERCC))
f.close()
print(missERCC)
```