Supporting Information for:

Accurate detection of chemical modifications in RNA by mutational profiling (MaP) with ShapeMapper 2

Steven Busan and Kevin M. Weeks*

Department of Chemistry, University of North Carolina, Chapel Hill NC 27599-3290

* correspondence: weeks@unc.edu

Six supporting figures.

SUPPORTING FIGURES

Figure S1. Multinucleotide mutation handling. Optimization of mutation separation threshold. ShapeMapper 2 merges nearby mutations and treats them as arising from a single inferred adduct. For two mutations to be taken as distinct, they must be separated by at least as many unchanged reference sequence nucleotides as specified by the `--min-mutation-separation` parameter; the default value is 6. For each separation threshold, area under the ROC curve was calculated over SHAPE reactivity values for the subset of A, U, G, or C nucleotide positions within the *E. coli* ribosomal RNA dataset. SHAPE reactivity profiles were evaluated relative to the *E. coli* ribosome structure from the Comparative RNA Web Site (Cannone et al. 2002).

Figure S2. Ambiguously aligned mutation handling. (A) Mechanistic model for the source of ambiguously aligned mutations. Ambiguously aligned mutations often appear to result from partial dissociation and reannealing of cDNA and RNA during reverse transcription. This model suggests that, during reverse transcription, cDNA and RNA mis-anneal through partial end complementarity thereby introducing deletions or insertions in an ambiguous local sequence context. This model implies that alignment to the 5' side of ambiguous mutations will more accurately recover adduct locations (blue ovals) than will alignment to the 3' side. (B) Empirical evaluation of 5' side versus 3' side realignment of ambiguously located mutations using the *E. coli* ribosomal RNA dataset. This analysis indicates that deletions and insertions more accurately recover base pairing information when aligned to the 5' side rather than the 3' side of the deletion or insertion. For this analysis, mutation profiles were created using only ambiguously aligned mutations. Mutation profiles were evaluated against the *E. coli* ribosome structure from the Comparative RNA Web Site (Cannone et al. 2002).

Figure S3. Effect of windowed read trimming and post-alignment basecall quality filter. (A) Read coverage is improved by windowed read trimming. Data are from an mRNA amplified with targeted RT-PCR. For hard trimming (black line), each read was scanned in the 5´ to 3´ direction, and downstream basecalls were discarded at the first basecall site not meeting a minimum

Phred quality score of 30 (estimated probability of incorrect basecall 0.1%). For windowed trimming (dashed blue line), downstream basecalls were discarded once a window of five nucleotides had an average quality score below 30. Note that windowed trimming will allow inclusion of some low-quality, isolated basecalls (yielding spurious mutations), necessitating a post-alignment basecall quality filter. Effective read depths after application of this filter are shown with a solid blue line. (B) Mutation rates calculated using window-trimmed reads without a post-alignment basecall quality filter. Note the high background rates around position 200. (C) Mutation rates calculated after applying a post-alignment basecall quality filter. This filter was implemented for both read depth and mutation rate as follows: Basecalls were excluded from contributing to the effective read depth if they or their immediate neighboring basecalls had a quality score below 30. Mutations were excluded from contributing to the mutation rate if they contained or were neighbored by basecalls with quality scores below 30.

Figure S4. Accuracy of ShapeMapper and electrophoresis data for small RNAs. True positives and false positives are defined as in Fig. 2. Electrophoresis data were collected previously, and reactivity profiles were evaluated against accepted structure models as described (Hajdin et al. 2013).

Figure S5. Example ShapeMapper 2 reactivity profile output figure. These plots are instructive for visualizing chemical probing data and for analyzing and troubleshooting problematic experiments. Data shown are from an *E. coli* thiamine pyrophosphate (TPP) riboswitch probed under ligand-bound conditions described previously (Siegfried et al. 2014). Top panel: SHAPE reactivities as read out by MaP, and estimated standard errors shown as error bars. These error bars are relatively small, indicating a high level of confidence in this reactivity profile. Middle panel: mutation rate profiles for the experimental and control samples, with standard errors indicated as lighter shaded areas. Comparing the red and blue profiles reveals a mutation rate signal significantly above background. Bottom panel: read depth profiles for each sample. This was a directed primer experiment, and the relatively flat read depth profiles indicate a robust PCR and minimal or nonexistent off-target primer binding. Effective read depths are shown in lighter colors, and show the effects of multinucleotide mutation handling and the basecall

quality filter.

Figure S6. Effect of aligner choice and of a denatured control sample. (A) ROC curves based on reactivity profiles from the *E. coli* rRNA dataset aligned using `STAR` or `bowtie2`. Both aligners result in highly accurate profiles with nearly identical agreement with structure models. (B) ROC curves showing SHAPE-MaP reactivity profile accuracies calculated with and without dividing the background-subtracted mutation rates by the mutation rates from a denatured control. Reactivity profiles were evaluated against the *E. coli* ribosome structure from the Comparative RNA Web Site (Cannone et al. 2002).

References

Cannone JJ, Subramanian S, Schnare MN, Collett JR, D'Souza LM, Du Y, Feng B, Lin N, Madabusi LV, Müller KM, et al. 2002. The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics* **3**: 2.

Hajdin CE, Bellaousov S, Huggins W, Leonard CW, Mathews DH, Weeks KM. 2013. Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proc Natl Acad Sci* **110**: 5498–5503.

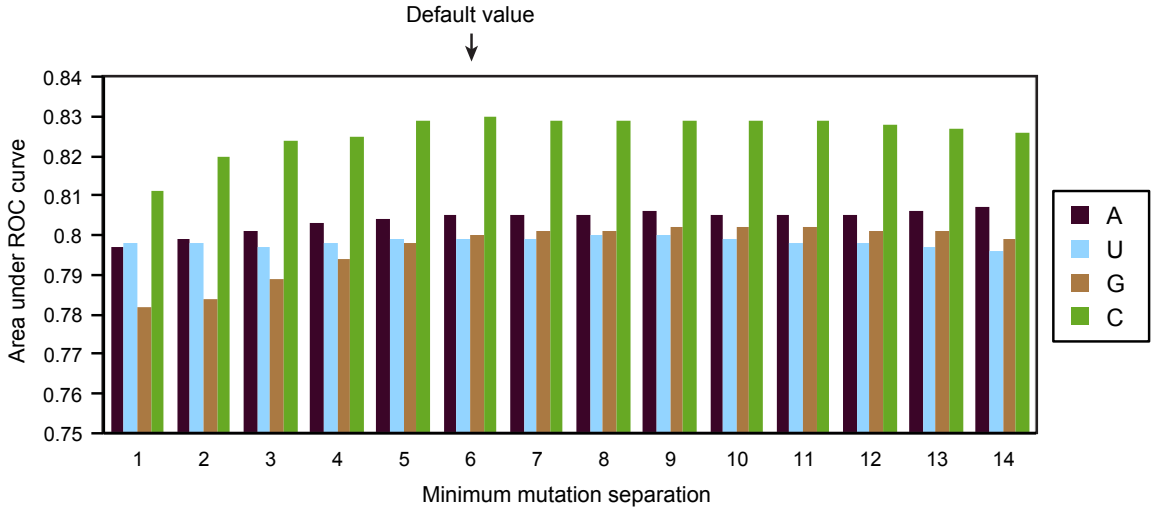
Default value
0.84
0.83
0.82
0.81
0.8
0.79
0.78
0.77
0.76
0.75
Area under ROC curve
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Minimum mutation separation
A
U
G
C

Figure S1

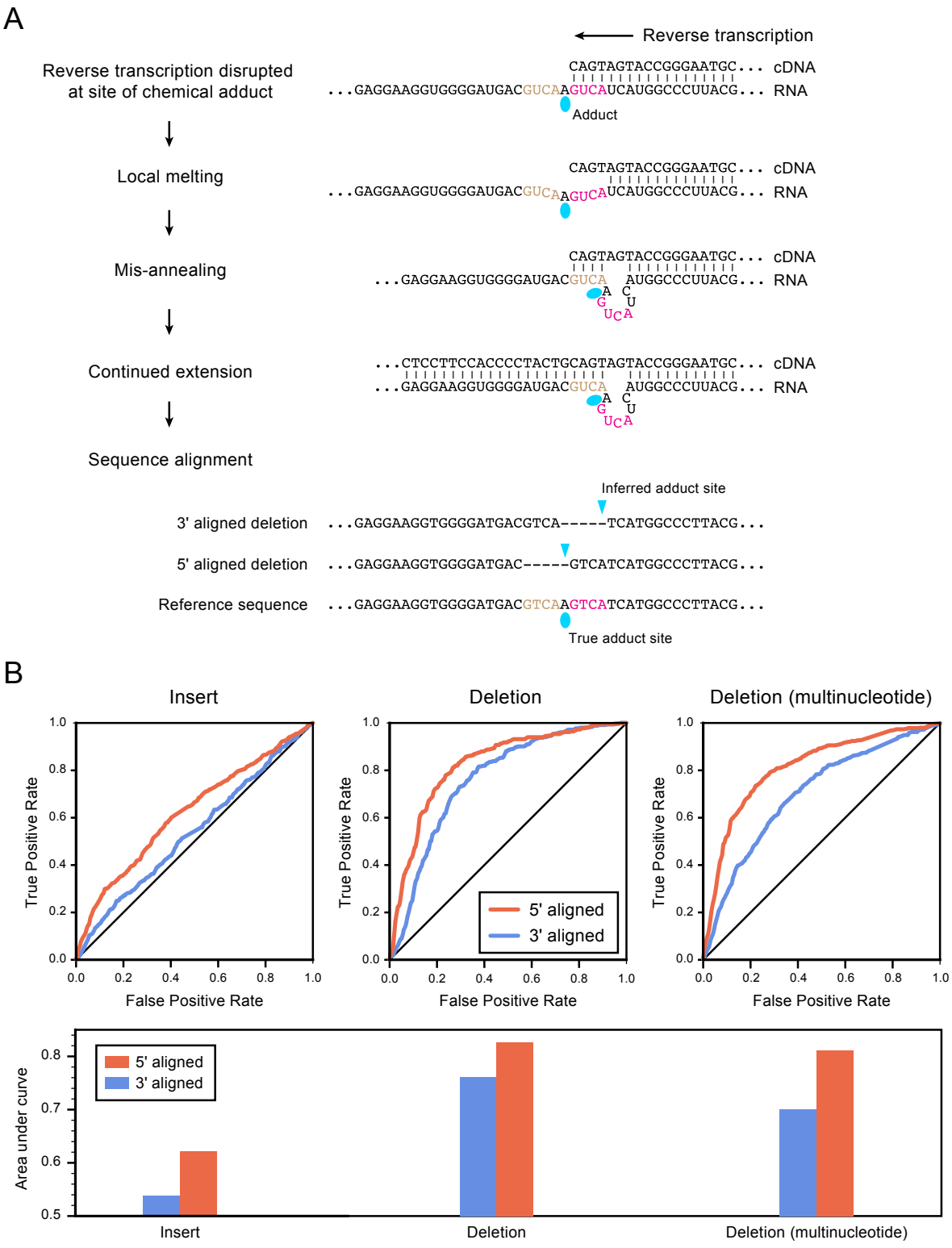
A
Reverse transcription
Reverse transcription disrupted at site of chemical adduct
Local melting
Mis-annealing
Continued extension
Sequence alignment
CAGTAGTACCGGGAATGC... cDNA
...GAGGAAGGUGGGGAUGACGUCAAGUCAUCAUGGCCCUUACG... RNA
Adduct
...CTCCTTCCACCCCTACTGCAGTAGTACCGGGAATGC... cDNA
...GAGGAAGGUGGGGAUGACGUCA AUGGCCCUUACG... RNA
Inferred adduct site
3' aligned deletion ...GAGGAAGGTGGGGATGACGTCA-----TCATGGCCCTTACG...
5' aligned deletion ...GAGGAAGGTGGGGATGAC-----GTCATCATGGCCCTTACG...
Reference sequence ...GAGGAAGGTGGGGATGACGTCAAGTCATCATGGCCCTTACG...
True adduct site
B
Insert
Deletion
Deletion (multinucleotide)
True Positive Rate
False Positive Rate
5' aligned
3' aligned
Area under curve
Insert
Deletion
Deletion (multinucleotide)

Figure S2

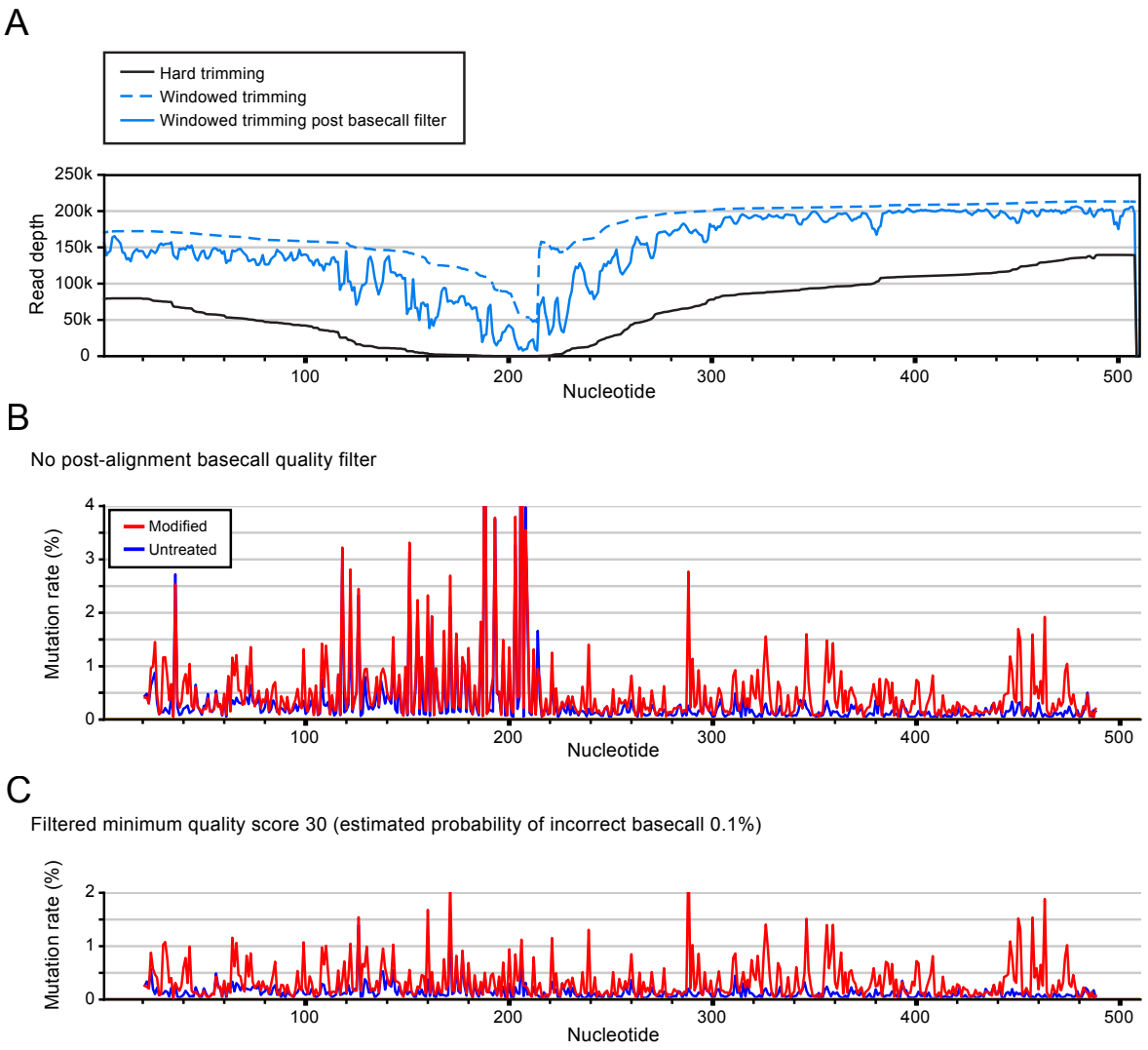
A
Hard trimming
Windowed trimming
Windowed trimming post basecall filter
Read depth
250k
200k
150k
100k
50k
0
100
200
300
400
500
Nucleotide
B
No post-alignment basecall quality filter
Modified
Untreated
Mutation rate (%)
4
3
2
1
0
100
200
300
400
500
Nucleotide
C
Filtered minimum quality score 30 (estimated probability of incorrect basecall 0.1%)
Mutation rate (%)
2
1
0
100
200
300
400
500
Nucleotide

Figure S3

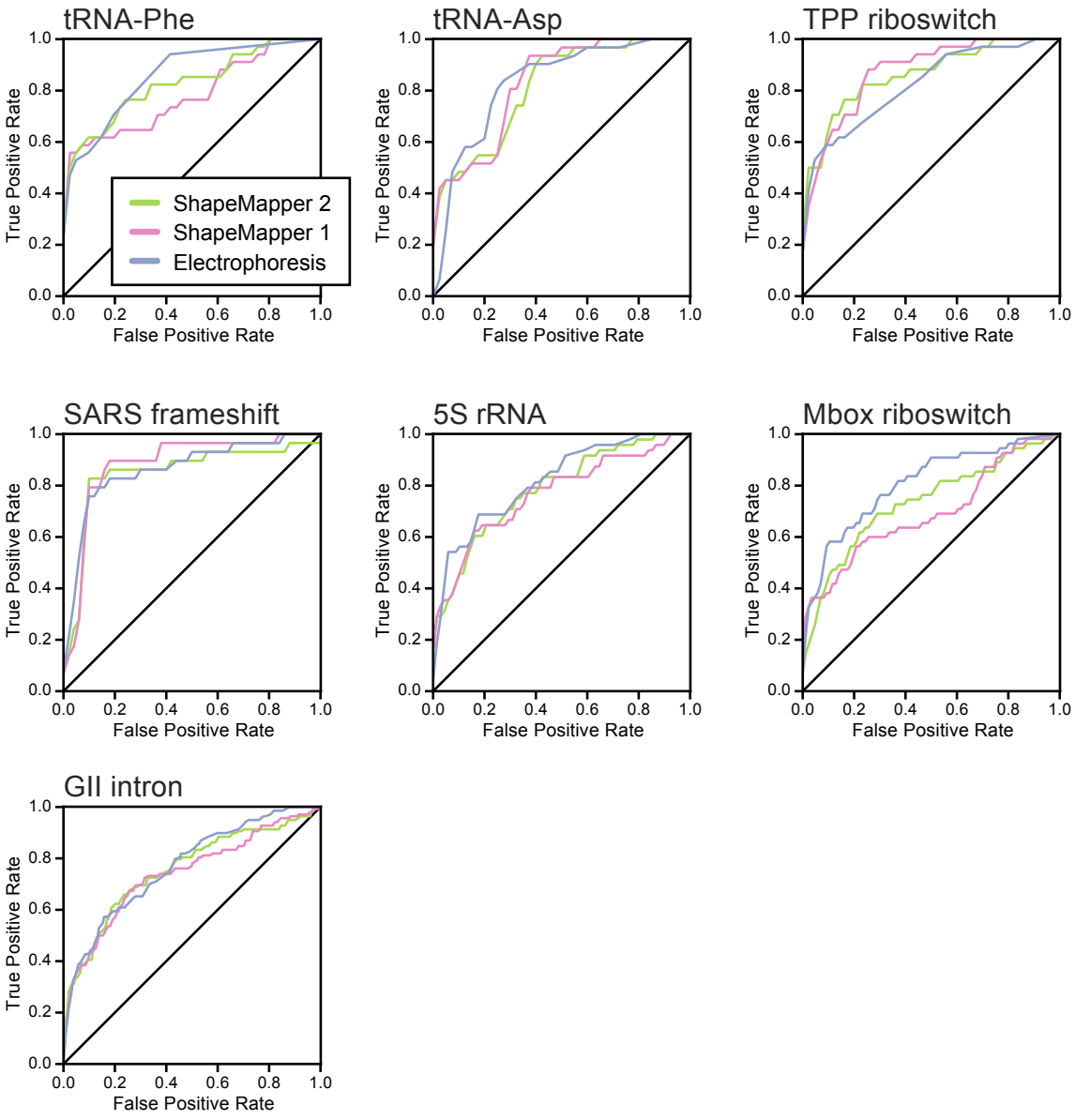
tRNA-Phe
tRNA-Asp
TPP riboswitch
SARS frameshift
5S rRNA
Mbox riboswitch
GII intron
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ShapeMapper 2
ShapeMapper 1
Electrophoresis

Figure S4

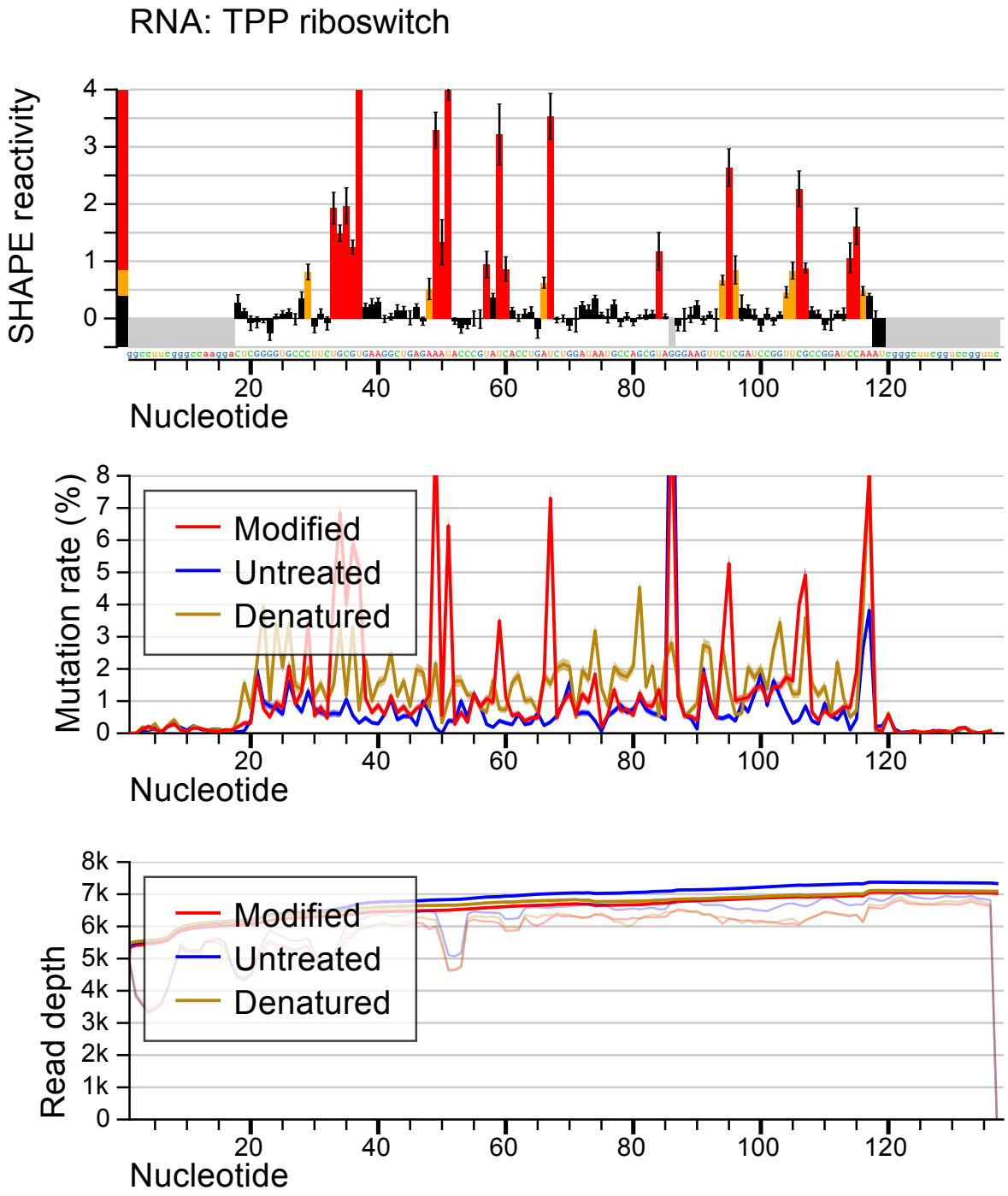
RNA: TPP riboswitch
SHAPE reactivity
Nucleotide
Mutation rate (%)
Modified
Untreated
Denatured
Nucleotide
Read depth
Modified
Untreated
Denatured
Nucleotide

Figure S5

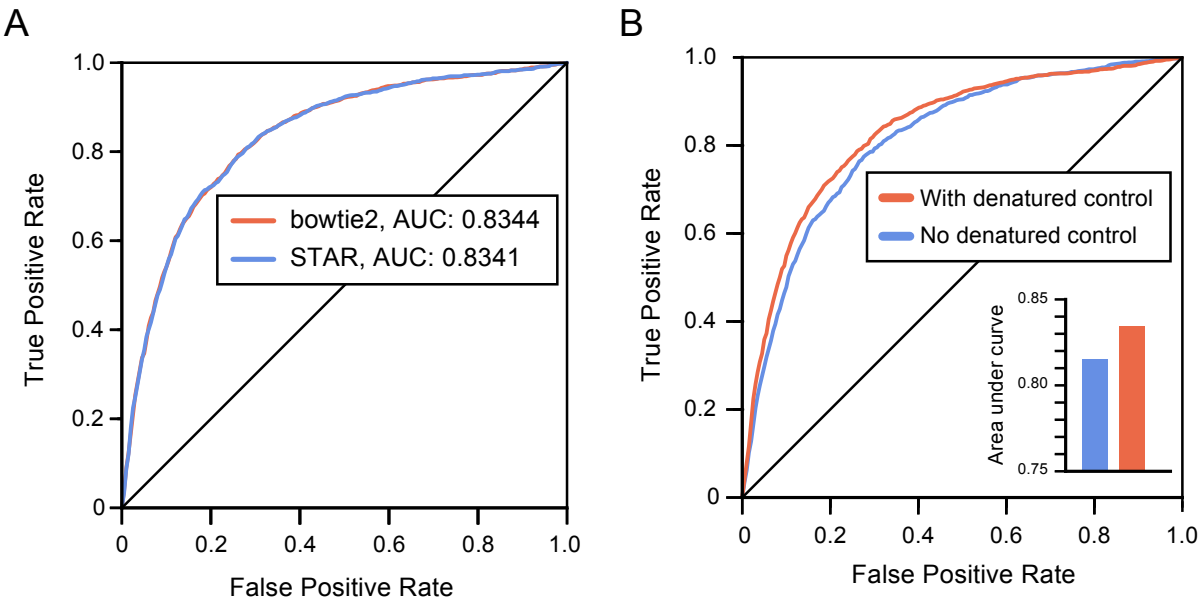
A
True Positive Rate
False Positive Rate
bowtie2, AUC: 0.8344
STAR, AUC: 0.8341
B
True Positive Rate
False Positive Rate
With denatured control
No denatured control
Area under curve

Figure S6