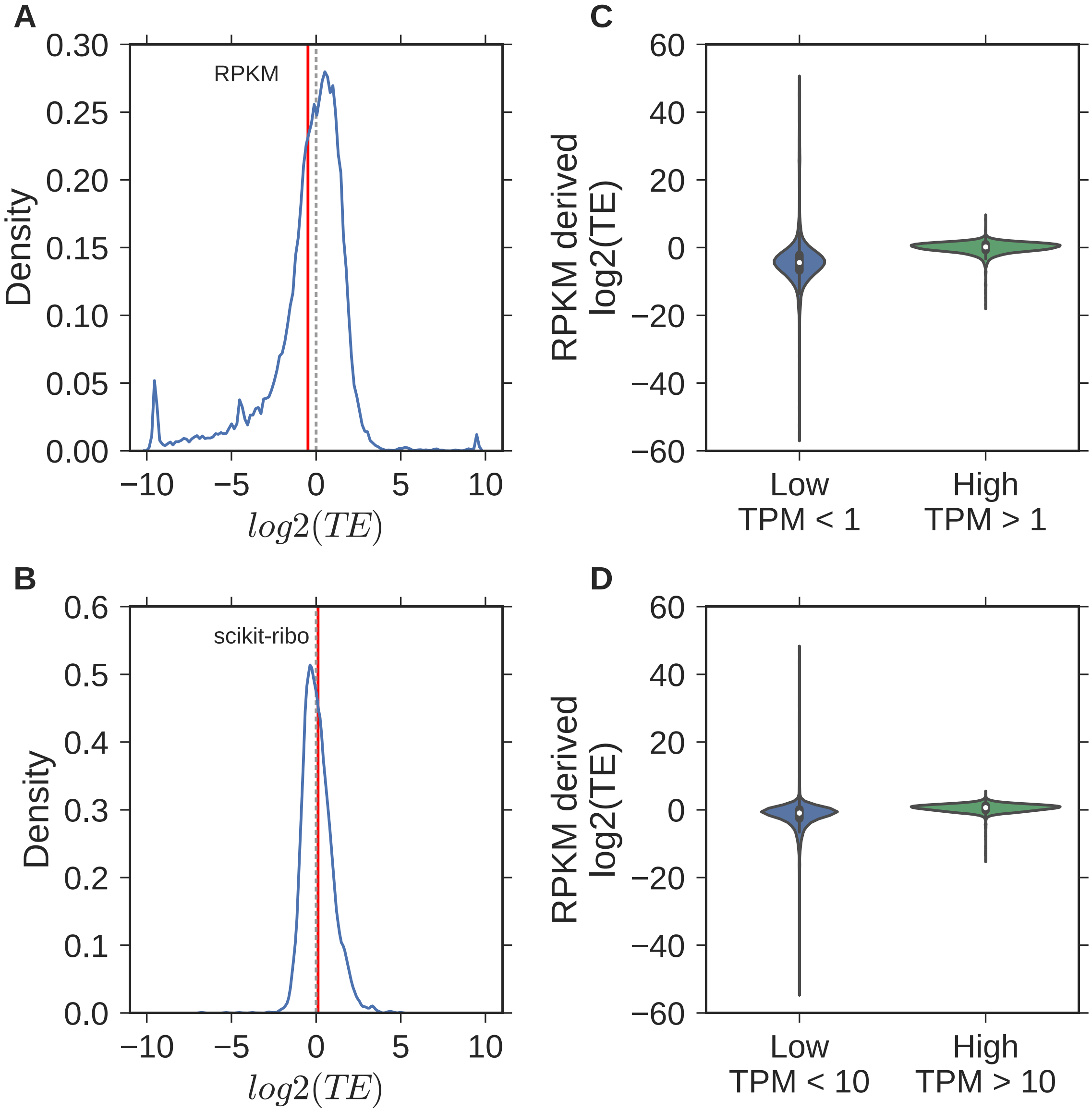
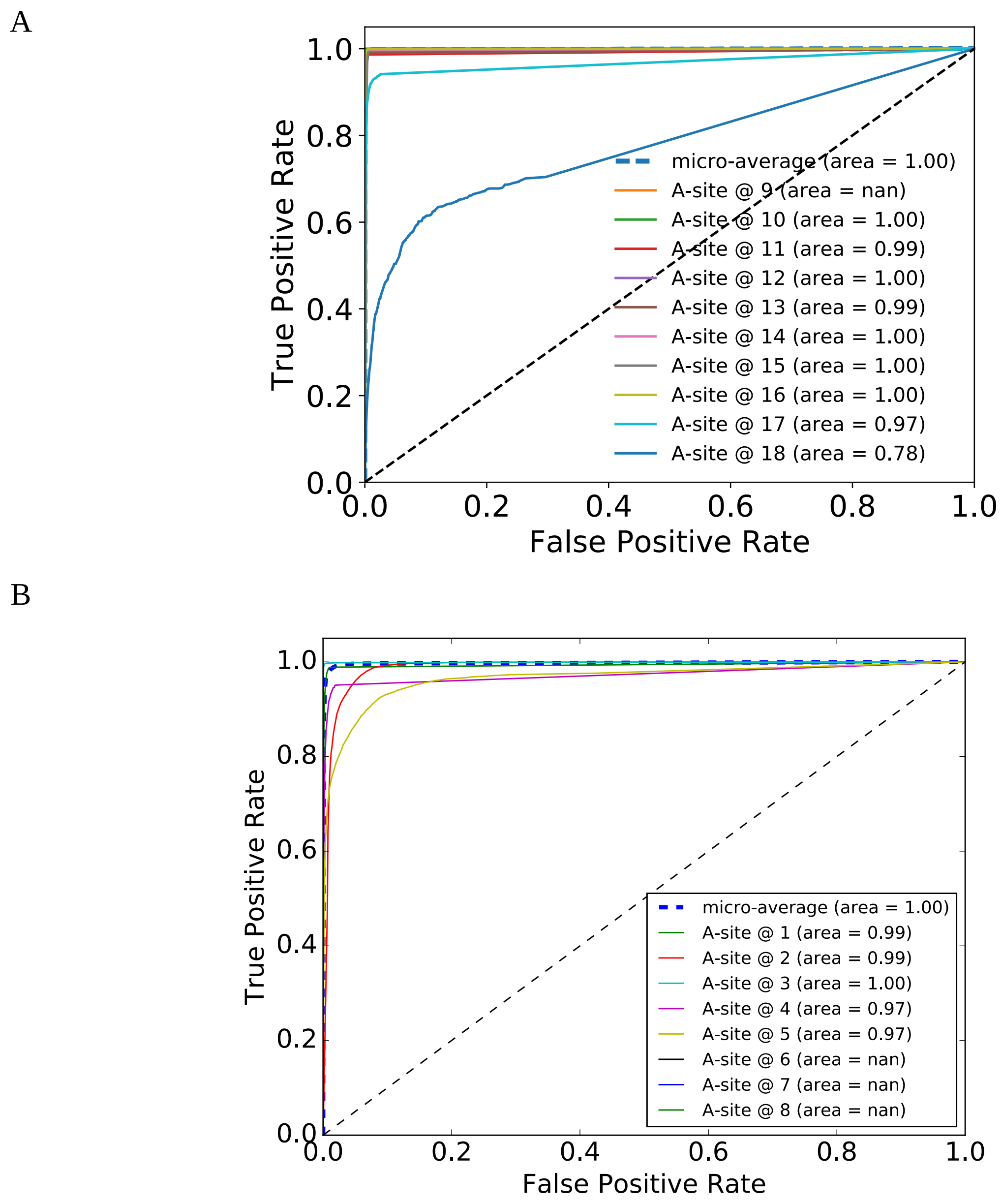
Supplemental information

Table of Contents

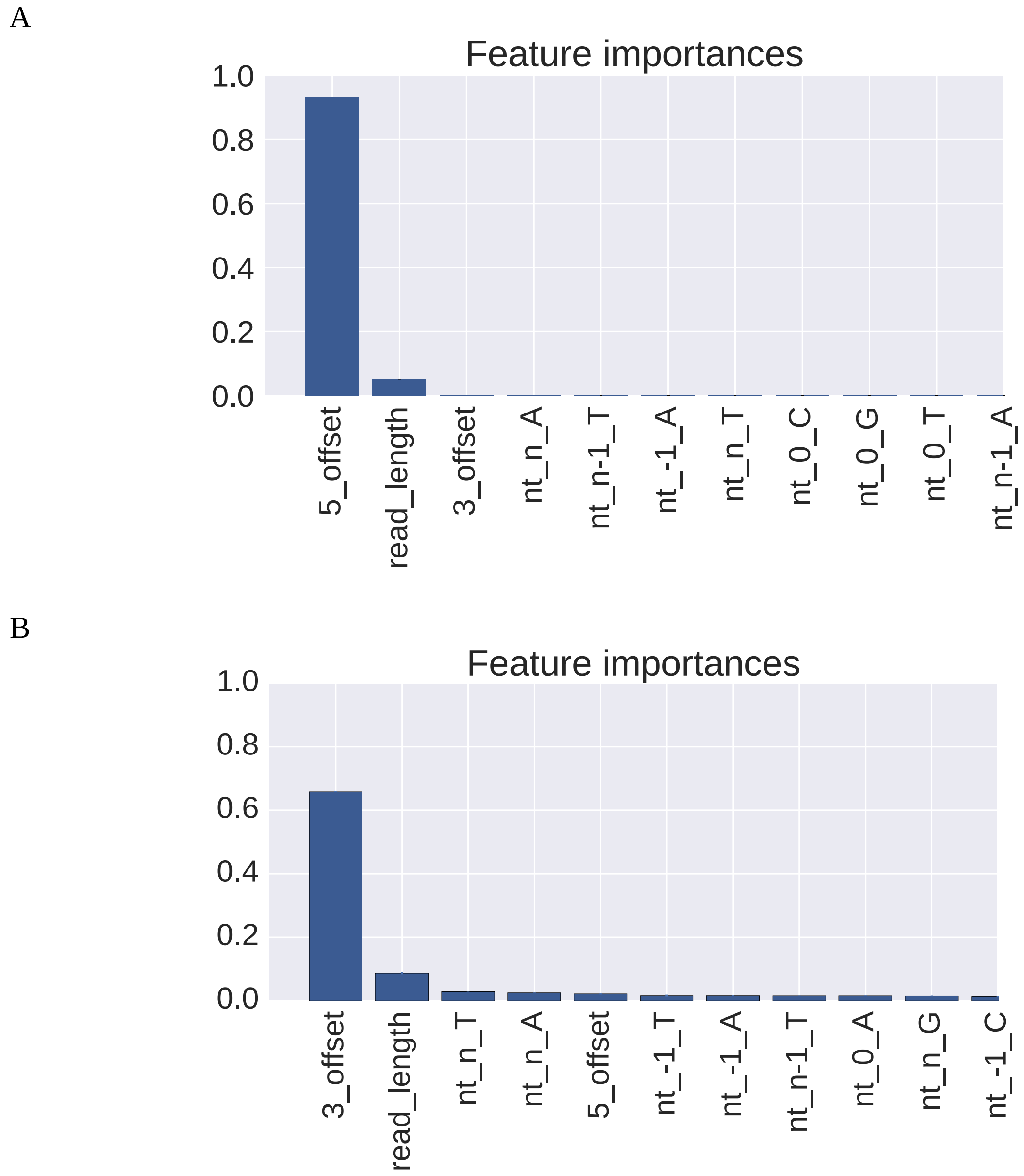
Supplemental figures



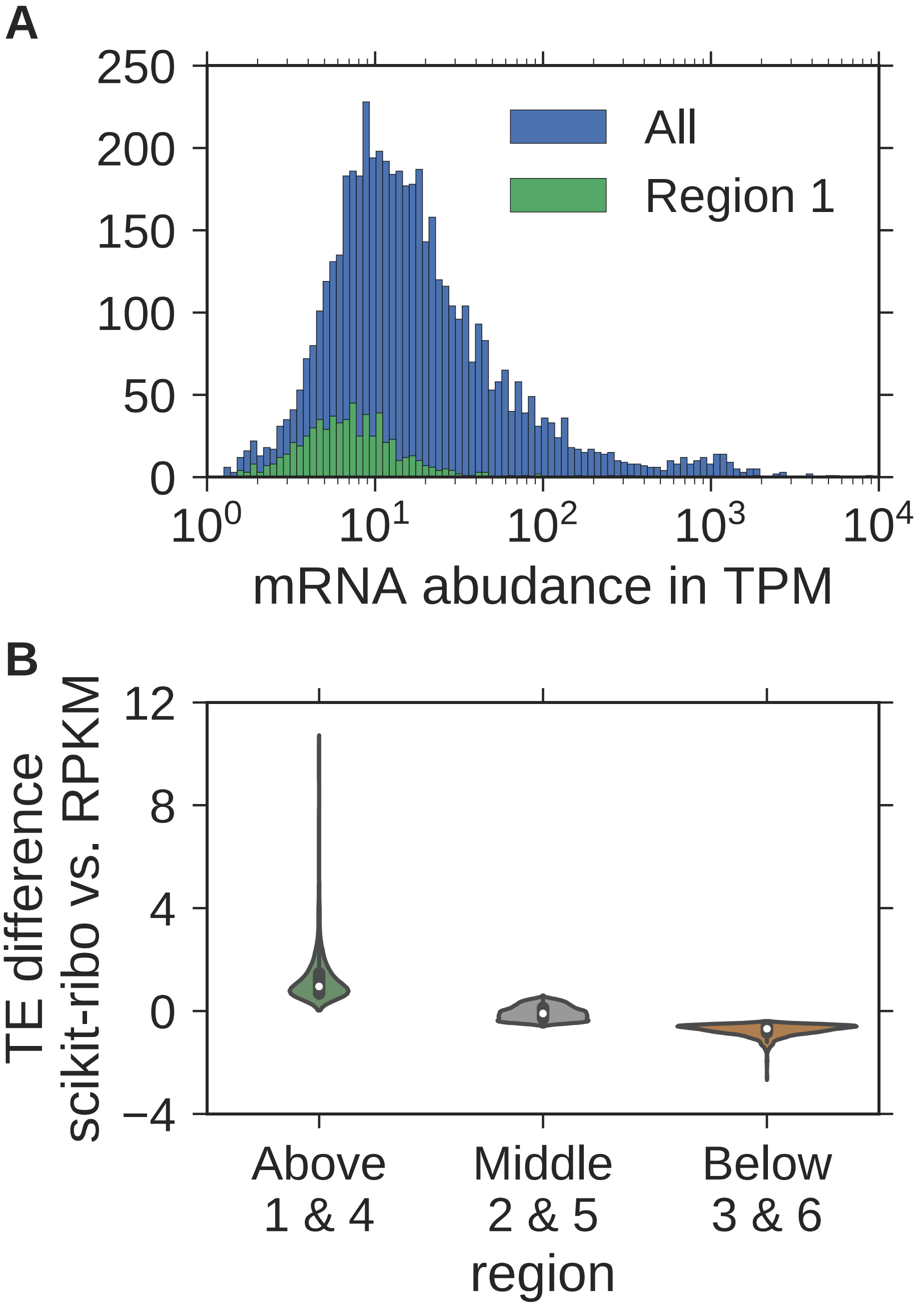
Supplemental figure S1. RPKM-derived log2(TE) and scikit-ribo log2(TE). Related to Figure 1. (A) The RPKM-derived log2(TE) reported high dispersion among low abundance genes (TPM<1), while the genes with TPM > 1 still reported a long tail on the negative side. (B) Scikit-ribo reported a balanced log2(TE) distribution (mean=0.1). The red solid line denotes the mean. (C) The RPKM-derived log2(TE) reported a skewed distribution (mean=-0.5). (D) Even increasing the TPM cutoff to 10, the RPKM-derived log2(TE) still reported a long tail on the negative side.



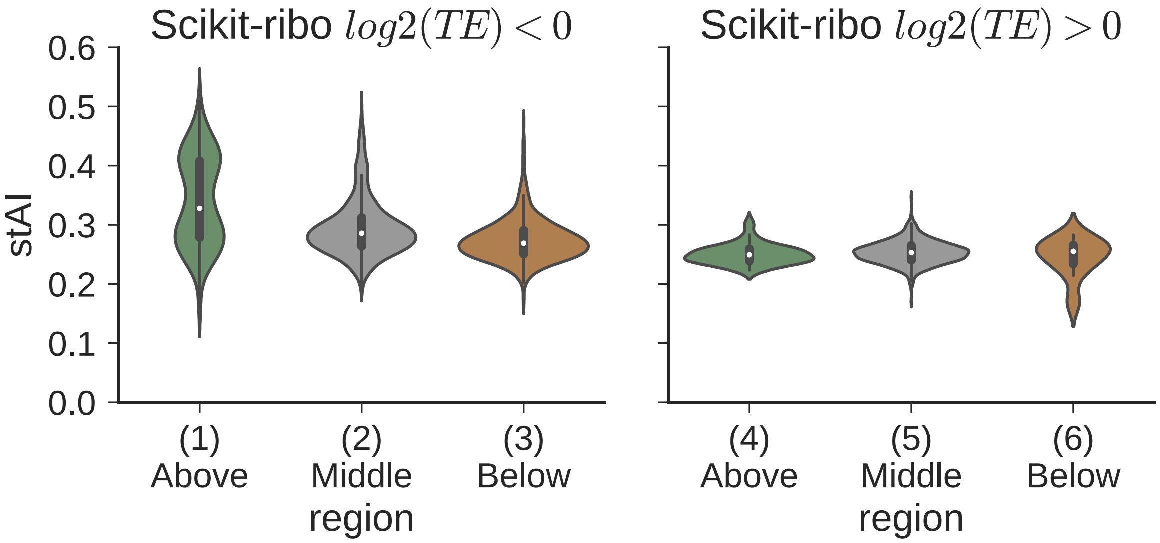
Supplemental figure S2. Multi-class ROC curves for A-site prediction. Related to Figure 2. (A) *S. cerevisiae* RNase I data. (B) *E. coli* RelE data. Each curve represents the data with different A-site locations (12 to 18 in RNase I, 1 to 8 in RelE). The dash line represents the micro-average across classes.



Supplemental figure S3. Feature importance from the random forest model. Related to Figure 2. (A) *S. cerevisiae* RNase I data. (B) *E. coli* RelE data. 5/3_offet represents whether the 5'/3' end of the read is in the first/second/third reading frame. Nt_-1/0/n-1/n represents the nucleotide at that position.

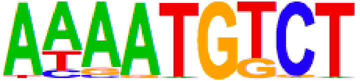
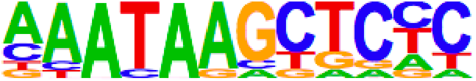
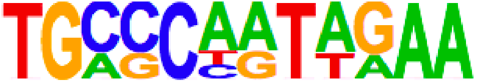
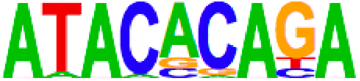
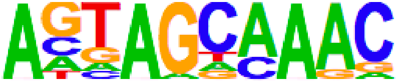
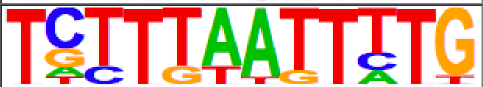
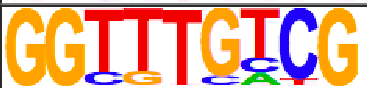
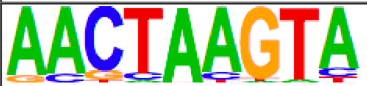
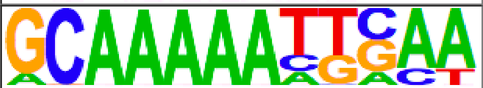
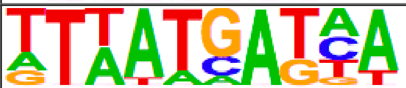
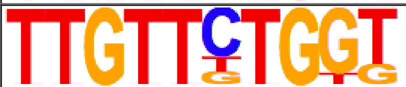
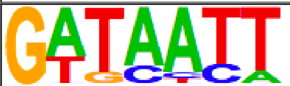
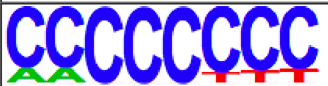
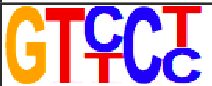


Supplemental figure S4. Analysis of mRNA abundance in TPM by region. Related to Figure 4; (A) Histograms of mRNA TPM in all genes (blue), and region 1 (green). (B) Violin plots of TE difference in the three regions, similar to Figure 4.



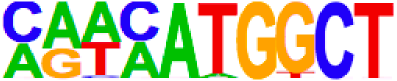
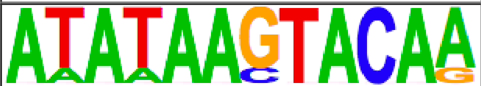
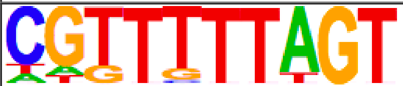
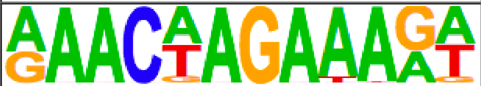
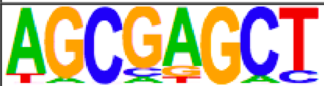
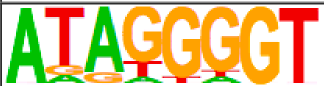
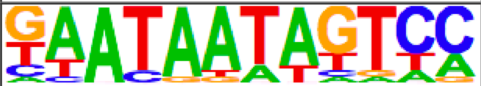
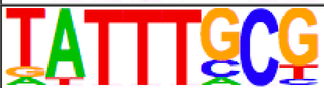
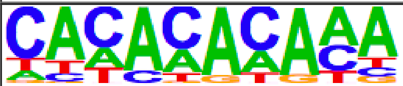
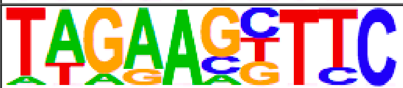
Supplemental figure S5. Violin plots of stAI for genes in the six regions. Related to Figure 4; left: $log2(TE) < 0$, right: $log2(TE) > 0$.

* - possible false positive

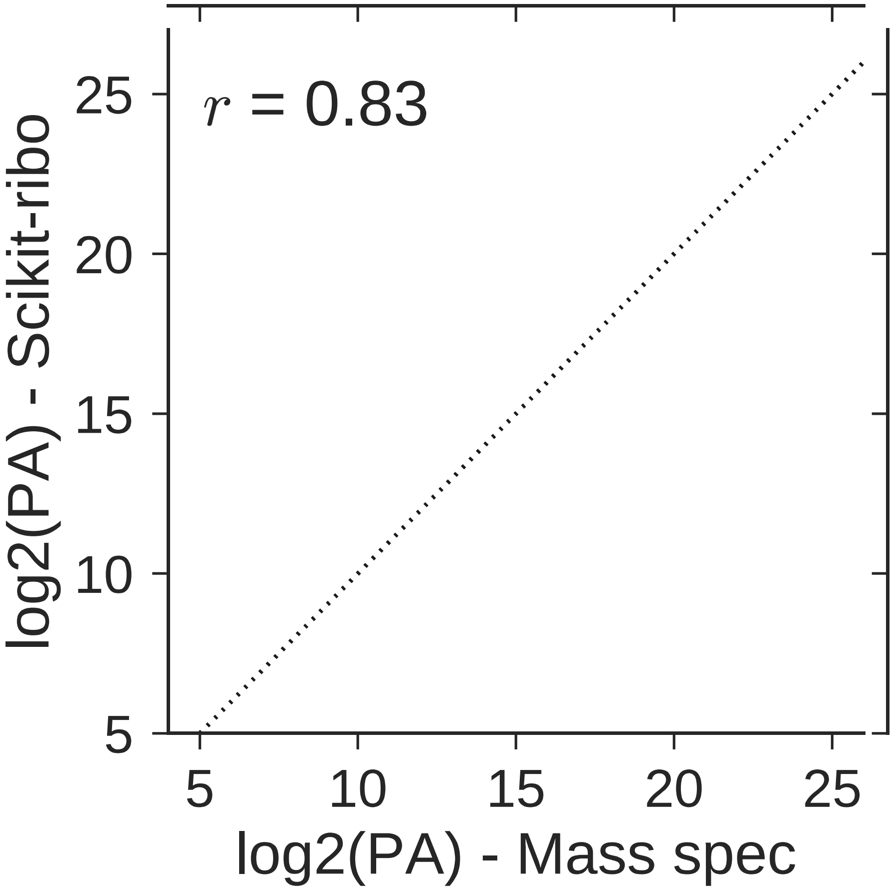
Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)
1		1e-21	-4.854e+01	21.09%	2.13%	29.2bp (31.9bp)
2 *		1e-11	-2.569e+01	13.61%	1.88%	14.6bp (16.2bp)
3 *		1e-10	-2.440e+01	4.08%	0.04%	8.7bp (14.6bp)
4 *		1e-9	-2.266e+01	14.29%	2.49%	28.6bp (36.0bp)
5 *		1e-8	-2.062e+01	5.44%	0.21%	11.6bp (15.4bp)
6 *		1e-8	-1.945e+01	5.44%	0.24%	9.2bp (16.0bp)
7 *		1e-8	-1.897e+01	5.44%	0.26%	20.4bp (31.5bp)
8 *		1e-8	-1.878e+01	16.33%	4.06%	27.9bp (38.8bp)
9 *		1e-7	-1.777e+01	4.08%	0.11%	6.1bp (16.6bp)
10 *		1e-6	-1.479e+01	7.48%	1.01%	17.3bp (17.7bp)
11 *		1e-6	-1.436e+01	2.72%	0.04%	9.5bp (11.8bp)
12 *		1e-4	-1.031e+01	25.17%	12.76%	28.9bp (35.5bp)
13 *		1e-1	-3.783e+00	0.68%	0.02%	0.5bp (15.8bp)
14 *		1e0	-2.205e+00	33.33%	28.41%	23.6bp (33.8bp)

Supplemental figure S6. Statistically enriched sequences based on scikit-ribo's TIE estimates using HOMER. Related to Figure 4; The Homer's suggested p-value threshold is 1×10^{-10} to 1×10^{-12}.

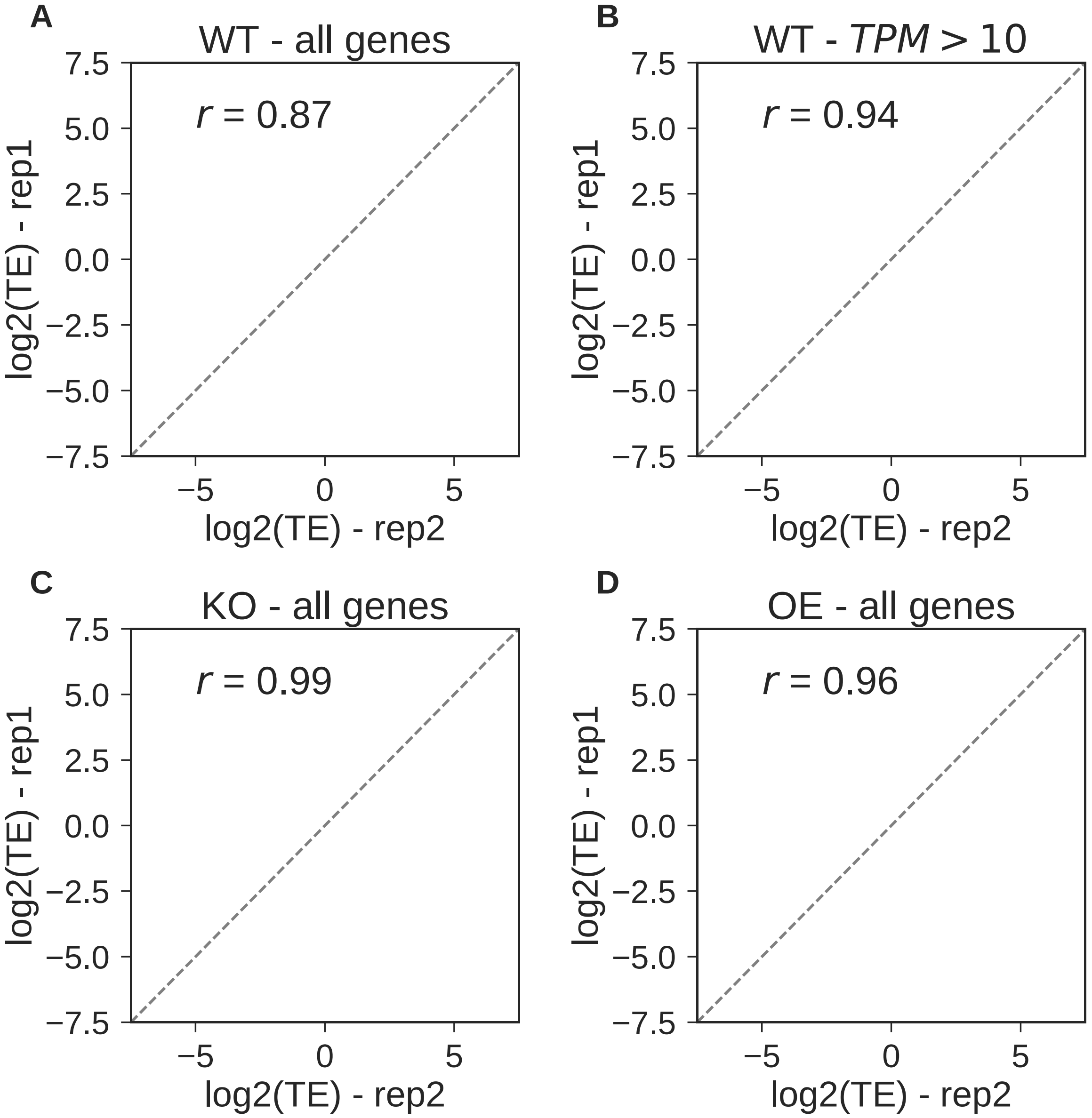
* - possible false positive

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)
1 *		1e-11	-2.587e+01	9.95%	1.34%	12.2bp (14.3bp)
2 *		1e-9	-2.123e+01	2.49%	0.02%	10.1bp (17.9bp)
3 *		1e-7	-1.829e+01	4.98%	0.40%	12.6bp (16.0bp)
4 *		1e-7	-1.705e+01	4.98%	0.45%	11.2bp (14.7bp)
5 *		1e-6	-1.586e+01	4.98%	0.52%	15.6bp (18.8bp)
6 *		1e-6	-1.442e+01	6.47%	1.12%	14.5bp (17.0bp)
7 *		1e-6	-1.393e+01	5.47%	0.80%	14.2bp (14.0bp)
8 *		1e-6	-1.385e+01	8.46%	2.03%	9.1bp (17.7bp)
9 *		1e-5	-1.199e+01	5.47%	0.98%	9.3bp (21.4bp)
10 *		1e-4	-1.111e+01	3.98%	0.53%	11.5bp (13.9bp)

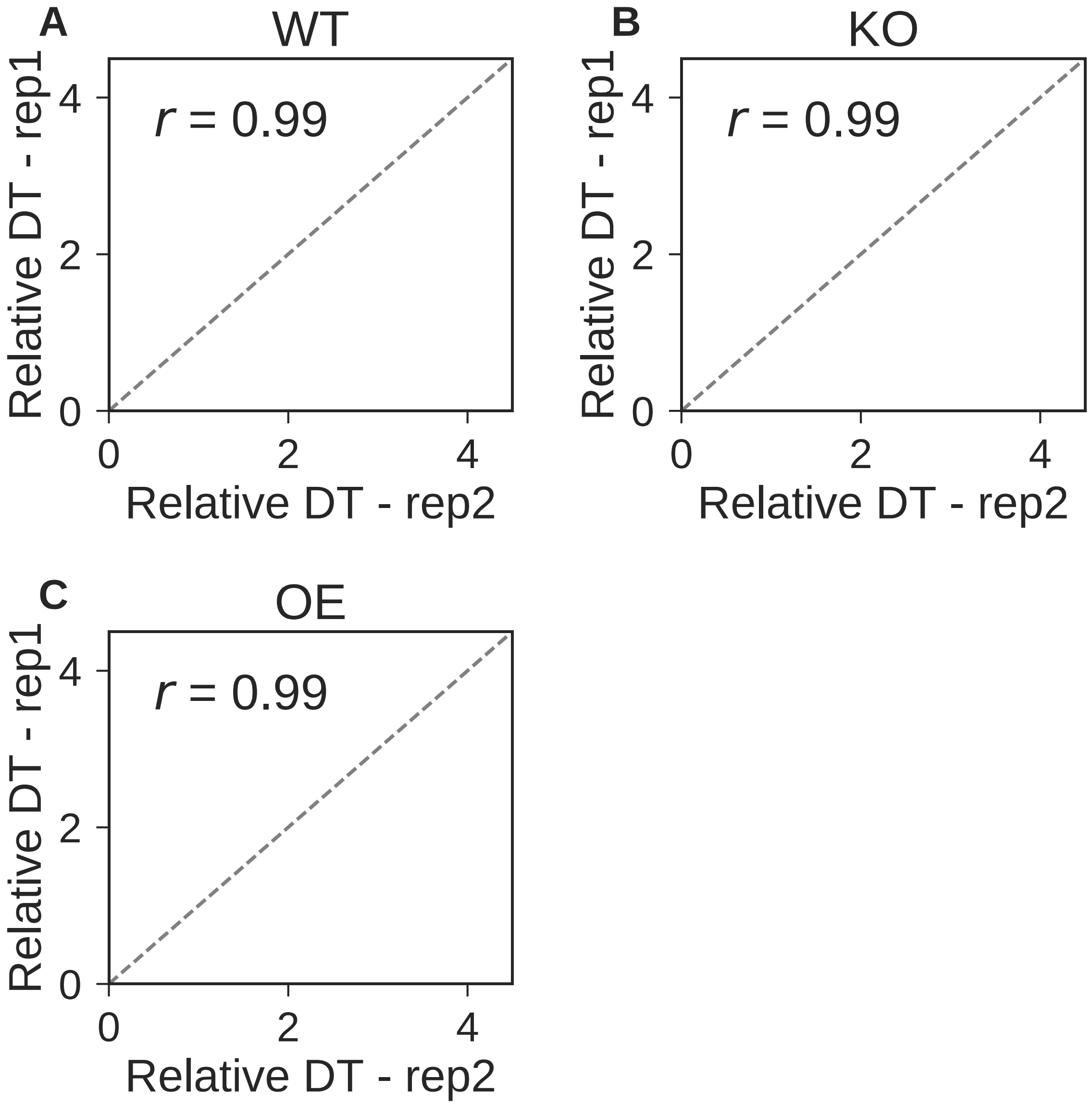
Supplemental figure S7. Statistically enriched sequences based on RPKM-derived TE estimates using HOMER. Related to Figure 4; The Homer's suggested p-value threshold is 1×10^{-10} to 1×10^{-12}.



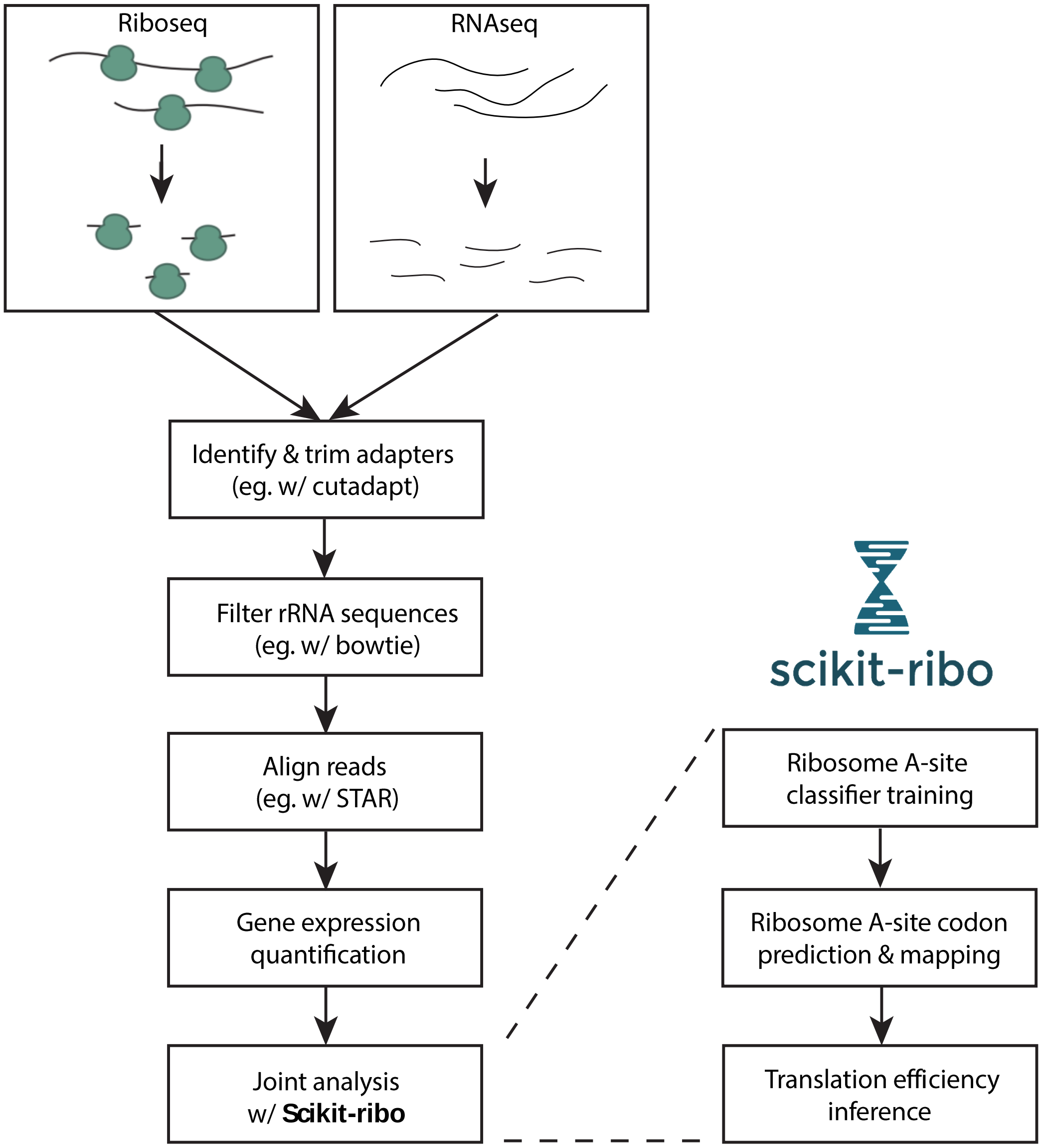
Supplemental figure S8. Higher correlation between scikit-ribo derived PA and SRM measurement, after considering protein degradation rate. Related to Figure 5; The protein degradation rate was obtained from Christiano et al ($r = 0.83$).



Supplemental figure S9. Highly reproducible TE estimates between replicates. Related to Figure 6; (A) WT: wild type, 55 million and 16.7 million in replicate 1 and 2 (r=0.87). (B) WT with TPM greater than (r=0.94). (C) KO: knock out *Dhh1p* (r=0.99), 74 million and 56 million in replicate 1 and 2. (D) OE: Overexpression of *Dhh1p*, 80 million and 39 million in replicate 1 and 2 (r=0.96). The correlation was a function of the number of reads in each replicate. the mean correlation of log(TE) were all very high between the biological replicates for a given strain (r=0.95), indicating that the data are of high quality and that the inference procedures in Scikit-ribo are stable.



Supplemental figure S10. High correlation of codon dwell time (DT) between biological replicates. Related to Figure 6; (A) wild-type, range of DT: 2.01, SD: 0.36, (B) KO, range: 3.05, SD: 0.45, (C) OE, range: 1.35, SD: 0.27. WT: wild type, KO: knock out *Dhh1p*, OE: Overexpression of *Dhh1p*. The mean correlation of relative DT and were all very high between the biological replicates for a given strain (r=0.99), indicating that the data are of high quality and that the inference procedures in Scikit-ribo are stable.



Supplemental figure S11. The complete workflow of Scikit-ribo analysis. First the RNA-seq and Riboseq sequencing reads are preprocessed to cut adapter sequences, filter rRNA reads, and then quantify the gene expression from the aligned RNAseq reads. After this pre-processing, Scikit-ribo is then used to predict the A-site locations and analyze the translation efficiency. Related to Figure 2.

Supplemental Tables

Study	SRR #	Mean accuracy	SD	# Optimal features
***S. cerevisiae* RNase I**				
Weinberg et al (2016)	SRR1049521	0.987	0.004	3
Radhakrishnan et al (2016)	SRR3493886	0.981	0.008	2
Radhakrishnan et al (2016)	SRR3493887	0.929	0.036	2
Radhakrishnan et al (2016)	SRR3493890	0.982	0.008	4
Radhakrishnan et al (2016)	SRR3493891	0.963	0.022	2
Radhakrishnan et al (2016)	SRR3493894	0.941	0.019	7
Radhakrishnan et al (2016)	SRR3493895	0.936	0.025	2
Radhakrishnan et al (2016)	SRR3493898	0.938	0.03	2
***E. coli* RelE**				
Hwang et al (2016)	SRR4023280	0.910	0.041	1
Hwang et al (2016)	SRR4023281	0.810	0.043	1

Supplemental Table S1. Prediction accuracy of A-site locations. Related to Figure 2. Mean and SD were computed via 10-fold cross validation. SD: standard deviation.

Region	Comparison	Sign of log2(TE)	# genes	Color
1	Under-estimated by RPKM	Negative	629	Green
2	Similar	Negative	1846	Gray
3	Over-estimated by RPKM	Negative	79	Orange
4	Under-estimated by RPKM	Positive	268	Green
5	Similar	Positive	1305	Gray
6	Over-estimated by RPKM	Positive	981	Orange

Supplemental table S2. Interpretation of the pair-wise comparison in Figure 4A. Related to Figure 4; The sign of log(TE) are based on TE of Scikit-ribo. $\Delta\, log2(TE) = log2(TE_{scikit-ribo}) - log2(TE_{RPKM})$. For gene with $\Delta\, log2(TE) < -0.5$, they were previously underestimated by RPKM-derived TE, and genes with $\Delta\, log2(TE) < -0.5$ were previously overestimated, and other genes have similar TE.

GO Term	Accession #	p-value	# genes
cytoplasmic translation	GO:0002181	3×10^{-25}	49
translational elongation	GO:0006414	1×10^{-8}	59
ribosome assembly	GO:0042255	2×10^{-6}	19
translation	GO:0006412	3×10^{-6}	63
peptide biosynthetic process	GO:0043043	4×10^{-6}	63

Supplemental Table S3. Gene set enrichment in region 4 genes. Related to Figure 4; There were 268 genes in region 4: 1) positive Scikit-ribo log2(TE), 2) previously under-estimated by RPKM derived TE. The p-values shown were adjusted with Bonferroni correction.