

## Effects of codon optimization on coagulation factor IX translation and structure:

### Implications for protein and gene therapies

Aikaterini Alexaki<sup>aa</sup>, Gaya K. Hettiarachchi<sup>aa\*</sup>, John C. Athey<sup>a</sup>, Upendra K. Katneni<sup>a</sup>, Vijaya Simhadri<sup>a</sup>, Nobuko Hamasaki-Katagiri<sup>a</sup>, Puja Nanavaty<sup>b</sup>, Brian Lin<sup>a</sup>, Kazuyo Takeda<sup>a</sup>, Darón Freedberg<sup>a</sup>, Dougald Monroe<sup>c</sup>, Joseph R. McGill<sup>a</sup>, Robert Peters<sup>d</sup>, Jacob M. Kames<sup>a</sup>, David D. Holcomb<sup>a</sup>, Ryan C. Hunt<sup>a</sup>, Zuben E. Sauna<sup>a</sup>, Amy Gelinias<sup>e</sup>, Nebojsa Janjic<sup>e</sup>, Michael DiCuccio<sup>f</sup>, Haim Bar<sup>g</sup>, Anton A. Komar<sup>b</sup>, and Chava Kimchi-Sarfaty<sup>a</sup>

<sup>a</sup> Center for Biologics Evaluation and Research, Food and Drug Administration, Silver Spring, MD, USA

<sup>b</sup> Center for Gene Regulation in Health and Disease, Cleveland State University, Cleveland, OH, USA

<sup>c</sup> University of North Carolina at Chapel hill, Chapel hill, NC, USA

<sup>d</sup> Bioverativ, Cambridge, MA, USA

<sup>e</sup> SomaLogic, Inc., Boulder, CO, USA

<sup>f</sup> National Center of Biotechnology Information, National Institutes of Health, Bethesda, MD, USA

<sup>g</sup> Department of Statistics, University of Connecticut, Storrs, CT, USA

Corresponding author:

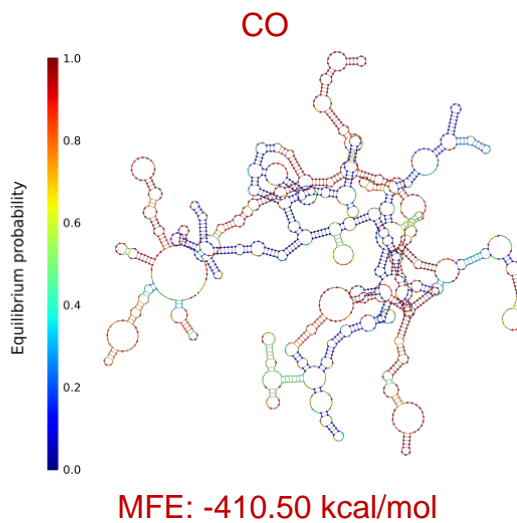
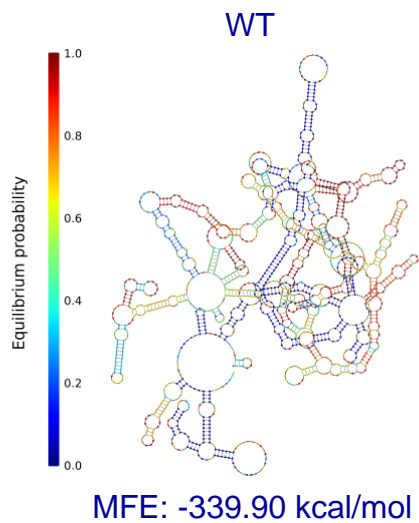
[Chava.kimchi-sarfaty@fda.hhs.gov](mailto:Chava.kimchi-sarfaty@fda.hhs.gov); Division of Plasma Protein Therapeutics, Office of Tissue and Advanced Therapies, Center for Biologics Evaluation and Research, Food and Drug Administration, Silver Spring, USA

\*Authors AA and GKH contributed equally

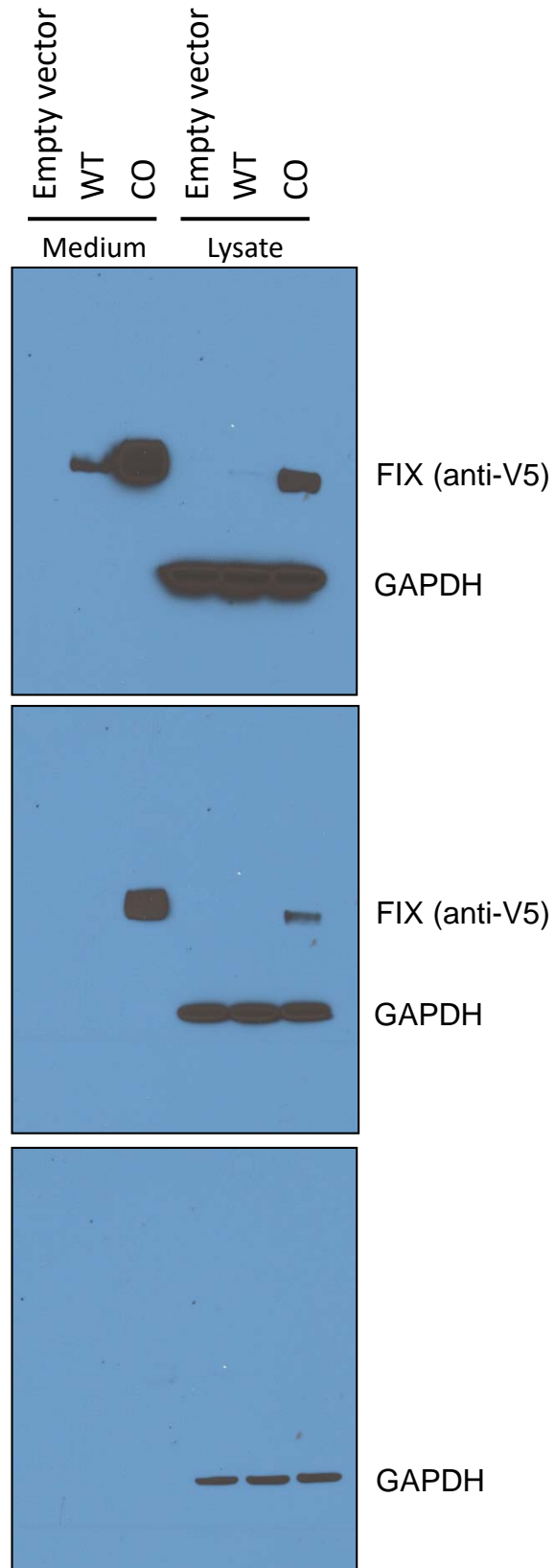
**Supplemental Figure S1. Comparison of the nucleotide sequences of codon-optimized and wild type *F9* variants.** Nucleotide changes between the two sequences are indicated in red, while codon changes are highlighted in yellow.

CO	1	ATGCAGAGAGTCAACAATGATATGGCTGAGTCACTGGGCTGATTACTAT	CO	749	CAITTTGGGGGAGGCTCTATCGTAAATGAGAAATGGATTGTACCGCCGCT	798
WT	1	ATGCAGCGGTGAACAATGATATGGCGAGATCAACCAGGCTCATCAACAT	WT	749	CAITTTGGGGGAGGCTCTATCGTAAATGAGAAATGGATTGTACCGGCTGCGC	798
		M Q R V N M I M A E S P G L I T I			F C G G S I V N E K W I V T A A	
CO	51	GTCCCTGCTGGGCTACTGCTGTCCCGCGAGTGTACCGGTTCCTGGGA	CO	799	CACGTGGTGGAAACCGGAGTCAASATCACAGTGTGCTGCTGGGAGCACAA	848
WT	51	GTCCCTTTTAGGATATCTACTACAGTGTGCTGAAATGTACAGTITTTCTTGA	WT	799	CACGTGTTTGAACCTGGTGTAAAAATACAGTITGTCGCAAGGTGAACATAA	848
		C L L G Y L L S A E C T V F L D			H C V E T G V K I T V V A G E H N	
CO	99	TCATGAGAACGCRAAATAAGATCTGAAACAGGCCAAAGATATAATAGTG	CO	849	CAITGAGGAAACAGACATACTGAGCAGAAAGCGAATGTGATCCGCATTA	898
WT	99	TCATGAAACGCCRAAATAATCTGAAATCGGCCAAAGAGTATAATTCAG	WT	849	TAITGAGGASACAGACATACTAGAGCAAAGCGAATGTGATTCGAAATTA	898
		H E N A N K I L N R P K R Y N S G			I E E T E H T E Q K R N V I R I I	
CO	149	GGAACTGGAGGAAATTTGTGACGGGCAACCTGGAGAGAGATGCATGGAG	CO	899	TTCCACACCACTAACTACAATGCAGCCATCAACAATAATAATCATGACATT	948
WT	149	GTAAATGGAGAGATTTGTCAAGGGAACCTGGAGAGAGATGTATGGAA	WT	899	TTCCACCCCAACTACAATGCAGCTATTAATAAGTACAAACATGACATT	948
		K L E E F V Q G N L E R E C M E			P H H N Y N A A I N K Y N H D I	
CO	199	GAAAAGTGTAGCTTCGAGGAAAGCCGCGAGAGTGTITGAAAATACAGAGCG	CO	949	GCCCTGCTGGAACTGGATGAGCCCTCTGGTGTCTAACAGCTACCTCACTCC	998
WT	199	GAAAAGTGTAGTITTTGAGGAAAGCCAGGAGAGTITTTGAAAACACTGAGAG	WT	949	GCCCTTCGGAACCTGGAGAACCTTATGTGCTAACAGCTACCTTACACC	998
		E K C S F E E A R E V F E N T E R			A L L E L D E P L V L N S Y V T P	
CO	249	AACCACAGAGTITGGAAAGCAGTATGTGGACGGGATCAGTGCAGAGCA	CO	999	AATCTGCATTGCTGCACAAAGAGTATACCAAATCTTCCCTGAAGTITGGAT	1048
WT	249	AACAACAGAAATITGGAAAGCAGTATGTGATGGAGATCAGTGTGATCCA	WT	999	TAITTGCAATGCTGCACAAAGAGTATACCAAATCTTCCCTGAATITGGAT	1048
		T T E F W K Q Y V D G D Q C E S N			I C I A D K E Y T N I F L K F G S	
CO	299	ACCCCTGTCTTAATGGCCGAGTGTGCAAGAAGATCAACTCATAAGAA	CO	1049	CAGGTAAGTGGAGCGGCTGGGGAAAGAGTCTTCCCAAGGGCAGGAGC-TC	1097
WT	299	ATCCAAGTITTAATGGCCGAGTGTGCAAGGATGACATTAATTCCTATGAA	WT	1049	CTGGTAAGTGGAGCGGCTGGGGAAAGAGTCTTCCCAAAAGGGA-GATCAGC	1097
		P C L N G G S C K D D I N S Y E			G Y V S G W G R V F H K G R S A	
CO	349	TGTGGTGTCCCTTCGGTITGAAAGCAAAATTCGAGCTGGAAGTGTAC	CO	1098	CCTGGTGTCCAGTATCTCGAGTGGCTCTGGTGTATCGAGCTACCTGTCT	1147
WT	349	TGTGGTGTCCCTTTGGTITGAAAGCAAAAGTGTGAAATGATGATGAC	WT	1098	TTTAGTGTCCAGTATCTTAGAGTTCCTCTGTGTGATCGAGCCACTGTCT	1147
		C W C P F G Y L E G K N C E L D V T			L V L Q Y L R V P L V D R A T C L	
CO	399	ATGTAACATTAAGAATGGACGTGCGAGCAGTITTTGAAAACTCTGCCG	CO	1148	TGAGTCTTACCAAGTITTACATCTAACAAACATGTTCTGGCTGGGTTT	1197
WT	399	ATGTAACATTAAGAATGGCAGTGTGCGAGCAGTITTTGAAAAATAGTGTCTG	WT	1148	TTCGATCTTACCAAGTITTACATCTAACAAACATGTTCTGGCTGGGTTT	1197
		C N I K N G R C E Q F C K N S A D			R S T K F T I Y N N M F C A G F	
CO	449	ATAAATAGGTGGTGTGCAAGCTGTACTGAAAGATATCGCCTGGCTGAGAAC	CO	1198	CAAGAGGGAGGAGAGATCTGTGTCAGGGGATTTCTGGGGGCCACATGT	1247
WT	449	ATAAATAGGTGGTITGCTCCCTGTACTGAGGGATATCGACTGCAGAAAC	WT	1198	CATGAGGGAGGTAGAGATCTGTGTCAGGGAGATAGTGGGGGACCCCATGT	1247
		N K V V C S C T E G Y R L A E N			H E G G R D S C Q G D S G G P H V	
CO	499	CAGAAGTCTCTGGAACCCAGCAGTCCCTTCCCTGTGGGAGGTGAGCGT	CO	1248	GACAGAGGTGAAAGGACCAGTTCCTGACTGGCATATTTCCTGGGGAG	1297
WT	499	CAGAAGTCTGTGAAACCCAGCAGTCCCTTCCCTGTGGAGAGTITCTGTG	WT	1248	TACTGAAAGTGAAGGACCAGTTCCTGACTGGCAATATTAGCTGGGGTG	1297
		Q K S C E P A V P F P C G R V S V			T E V E G T S F L T G I I S W G E	
CO	549	CTCCAGACTTCAAACTGACAGAGCAGAGACAGTITTTCCCGACTGG	CO	1298	AGGAAATGTGCAATGAAGGGAATAACGGGATCTACACCAAGTGAAGCGC	1347
WT	549	ITCACAACTTCTAAGCTGACCGTGTGAGACTGTITTTCCCTGATGTGG	WT	1298	AAGAAATGTGCAATGAAGGCAATAATGGAAATATATACCAAGTATCCCGG	1347
		S Q T S K L T R A E T V F P D V D			E C A M K G K Y G I Y T K V S R	
CO	599	ATTACGTCAACAGCACTGAGGCCGAAACCATCTGGCAACATTACTCAG	CO	1348	TATGTAACTGGATCAAGGAAAAAACCAACTGACC	1383
WT	599	ACTAATGTAATTTCTACTGAAAGCTGAAACCATTTTGGATAACATCACTCAA	WT	1348	TATGTAACTGGATTAAAGAAAAAACCAAGCTCACT	1383
		Y V N S T E A E T I L D N I T Q			Y V N W I K E K T K L T	
CO	649	TCTACCCAGAGTITCAATGACTTACTCGGGTGGTGGGGGGAGGATGCG	CO	649	AGGACCCAAAGCAITTAATGACTTACTCGGGTITGTTGGTGGAGAGATGCG	
WT	649	AGGACCCAAAGCAITTAATGACTTACTCGGGTITGTTGGTGGAGAGATGCG	WT	649	S T Q S F N D F T R V V G G E D A	
		S T Q S F N D F T R V V G G E D A				
CO	699	TAAACCAAGCCAGTTCGCCCTGGCAGTGGTCTGAAAGGAAAGGTGATG	CO	699	TAAACCAAGCCAGTTCGCCCTGGCAGTGGTCTGAAAGGAAAGGTGATG	
WT	699	CAAACCAAGGTCAAATCCCTTGGCAGTGGTCTGAAAGGAAAGGTGATG	WT	699	CAAACCAAGGTCAAATCCCTTGGCAGTGGTCTGAAAGGAAAGGTGATG	
		K P G Q F P W Q V V L N G K V D A				

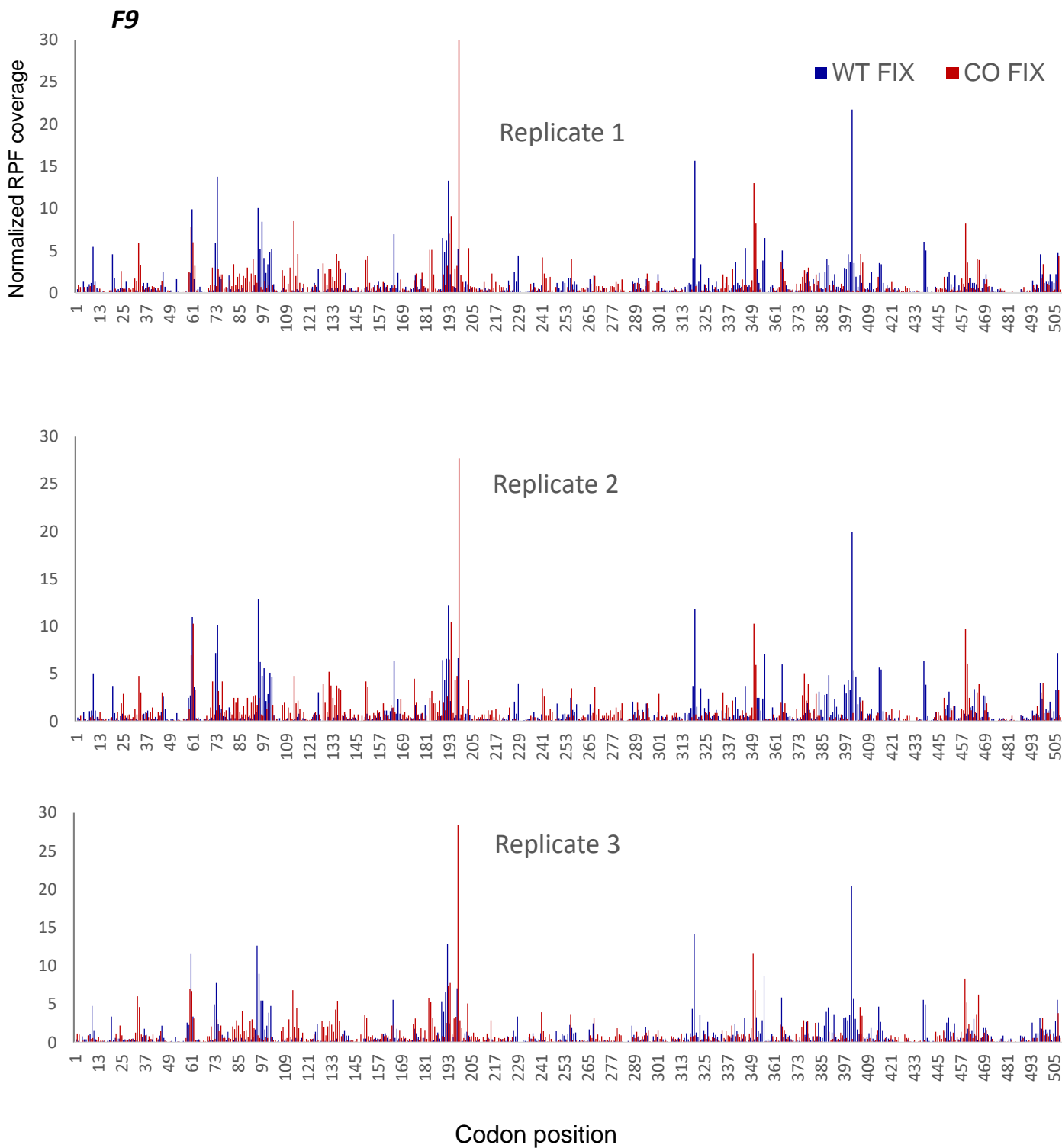
**Supplemental Figure S2. In silico analyses of mRNA folding and stability for the WT and CO F9 coding sequence.** Predicted MFE values and structures are shown for 37°C based on NUPACK algorithm ([www.nupack.org](http://www.nupack.org))



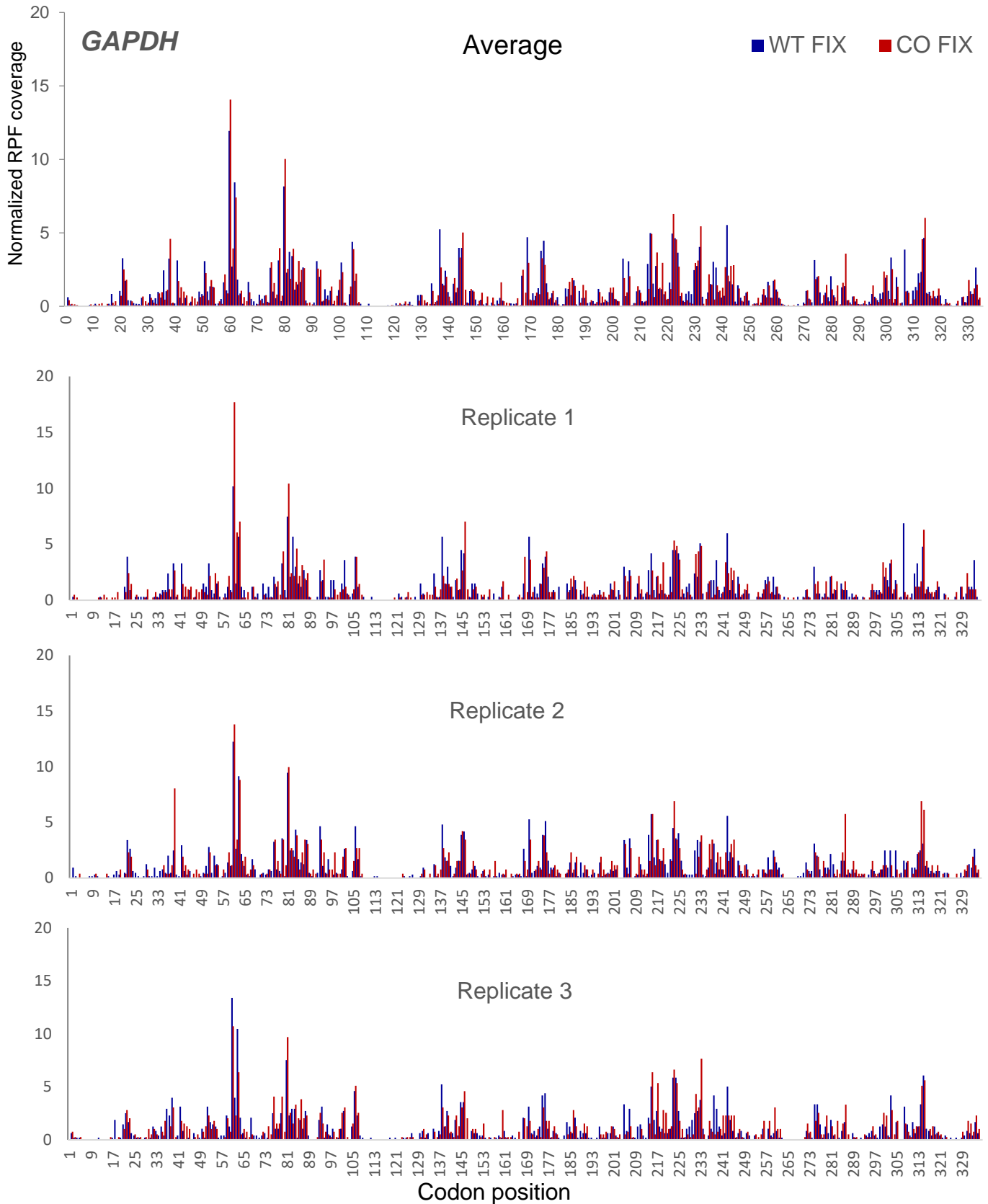
**Supplemental Figure S3. FIX expression is higher in HEK293T cells stably transfected with wild-type vs. codon-optimized F9.** Cell supernatant (concentrated medium) and lysate was collected and analyzed by immunoblotting. Membranes were stained with anti-V5 or anti-GAPDH. Multiple exposures of film are shown to demonstrate FIX (top image) and GAPDH (bottom image) at appropriate contrast. In the medium lanes, 5  $\mu$ g of protein were loaded per lane. In the lysate lanes, 10  $\mu$ g of protein were loaded per lane. Portions of these images were used to create Figure 2b.



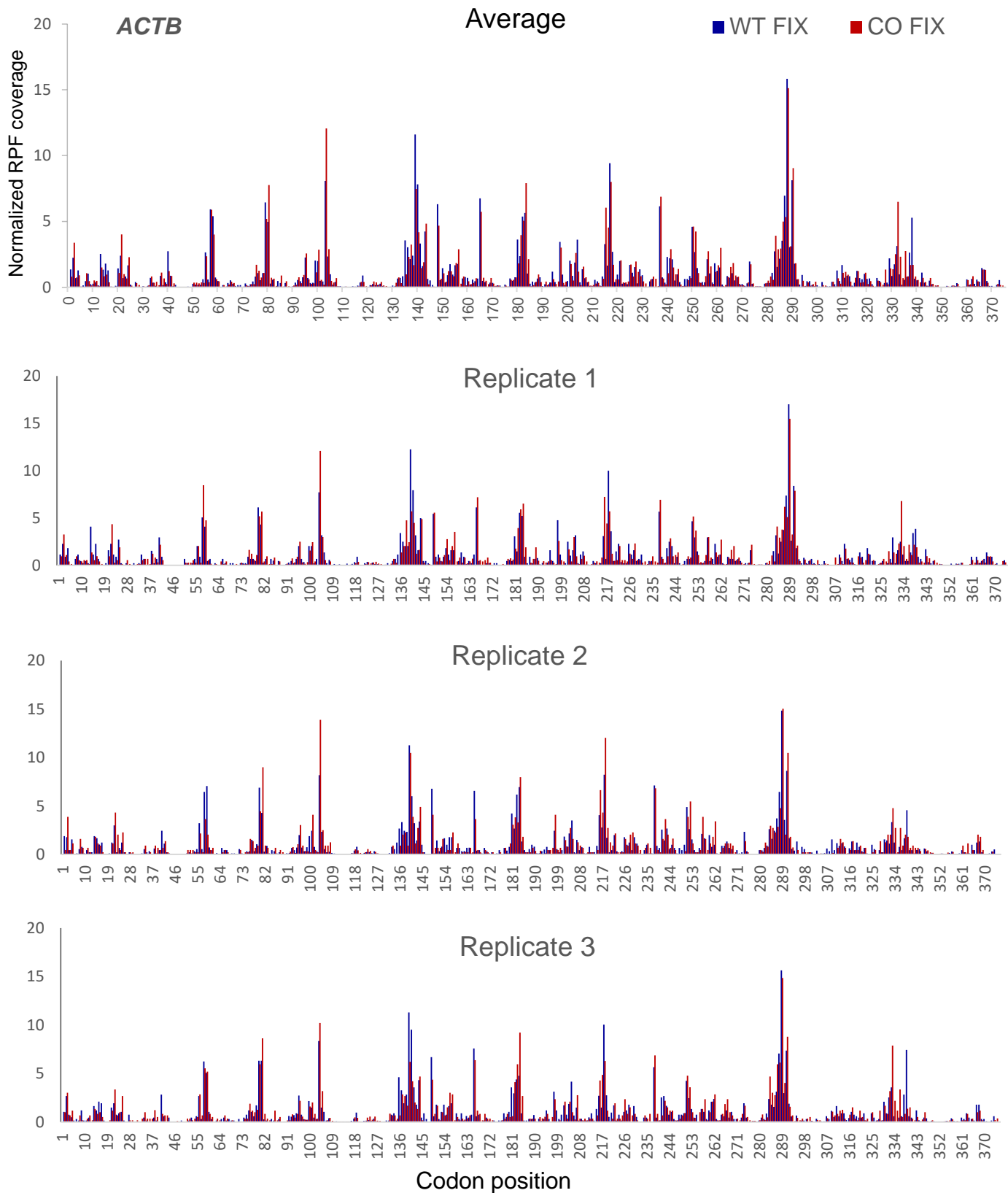
**Supplemental Figure S4.** RPF densities along the WT and CO *F9* mRNA for three replicates show reproducibility of results.



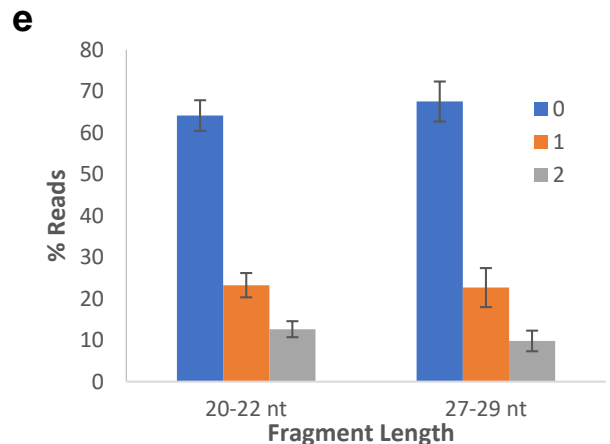
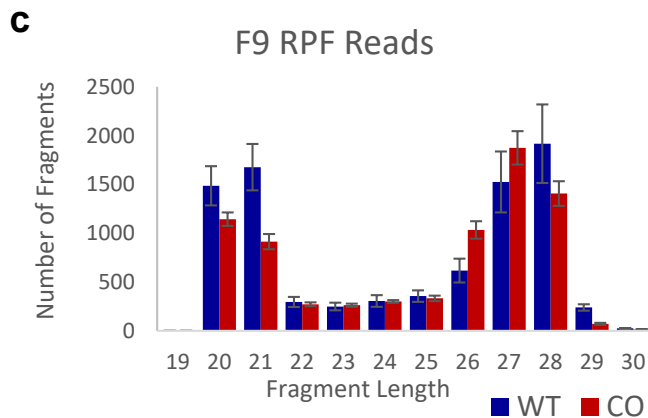
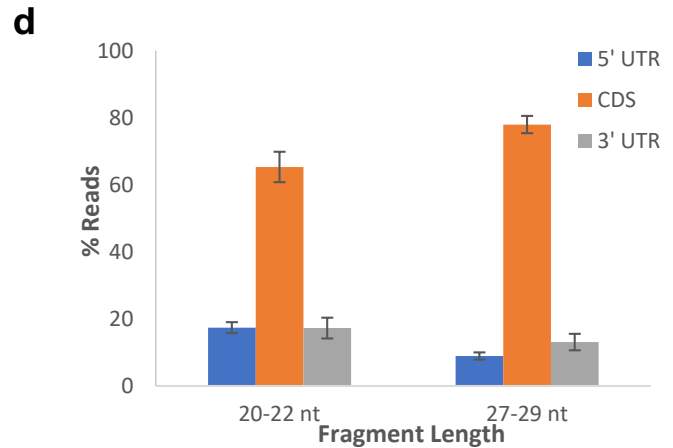
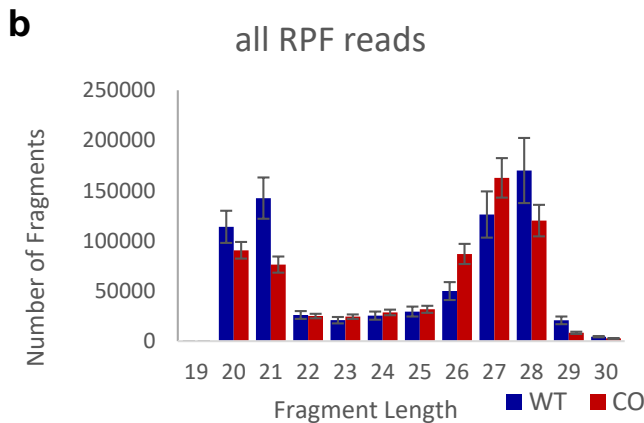
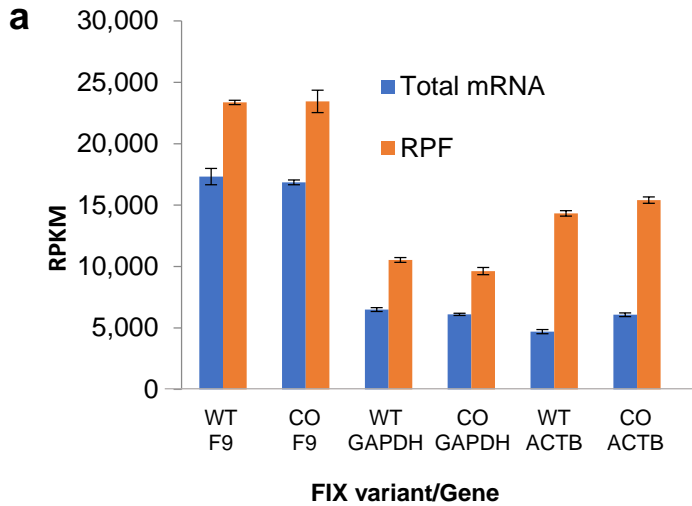
**Supplemental Figure S5. *GAPDH* is translated with very similar kinetics in the wild type and codon-optimized expressing cell lines.** RPF densities along the *GAPDH* mRNA are shown as average of 3 experiments and in each individual ribosome profiling experiment.



**Supplemental Figure S6. *ACTB* is translated with very similar kinetics in the wild type and codon-optimized expressing cell lines.** RPF densities along the *ACTB* mRNA are shown as average of 3 experiments and in each individual ribosome profiling experiment.

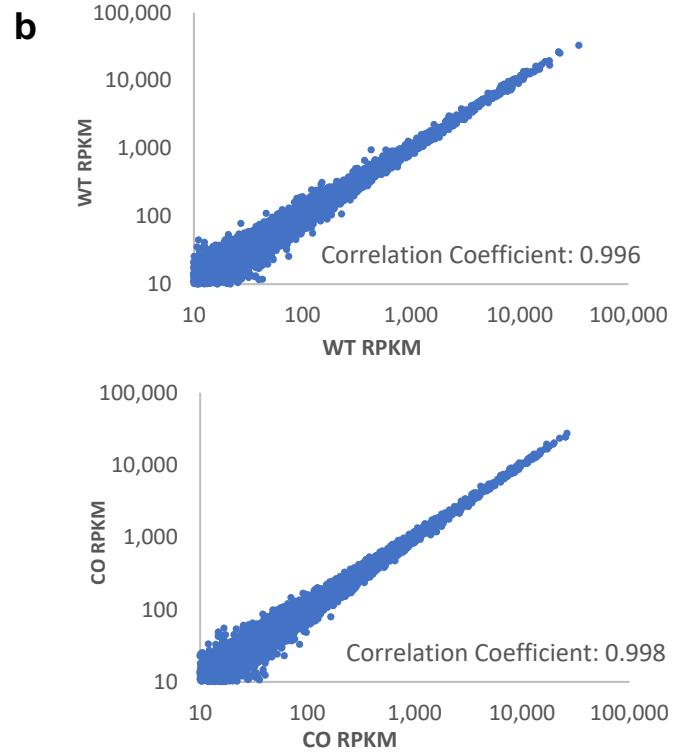
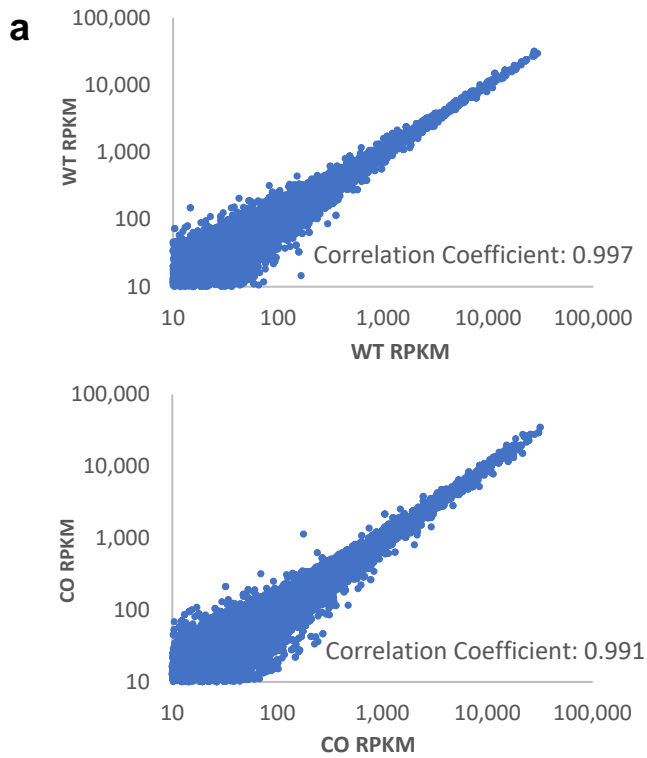


**Supplemental Figure S7. Multiple sets of controls show consistency in ribosome profiling experiments and analysis** (a) Transcript and RPF abundance for *F9*, *GAPDH* and *ACTB* in the WT and CO FIX expressing cell lines. The average of 3 experiments was plotted. (b) RPF size distribution of all reads in the WT and CO FIX expressing cell lines. (c) *F9* RPF size distribution in the WT and CO FIX expressing cell lines. (d) RPF sizes 20-22 and 27-29nt mapped 5'UTR, CDS and 3'UTR regions, (c) triplet periodicity of RPF fragment sizes 20-22 and 27-29 nucleotides, the average of 3 experiments was plotted, s.e.m. are shown.





**Supplemental Figure S8. Reproducibility of ribosome profiling experiments.** Pearson's correlation for (a) RPFs and (b) total mRNA between two representative experiments.



**Supplemental Table 1.** In a multiple linear regression model, the ribosome occupancy of adjacent codons is the only consistently significant predictor of ribosome occupancy of a codon.

	Estimate	Std. Error	t-value	Pr(> t )	Significance
<b>WT F9</b>					
(Intercept)	-1.05664	0.50251	-2.103	0.036	*
RSCU_Genome	0.03087	0.89716	0.034	0.9726	
RSCU_HEK_SRR	0.55596	1.06783	0.521	0.6028	
RSCU_HEK_ERX	-0.44406	1.23987	-0.358	0.7204	
RSCPU_Genome	-0.49032	0.30636	-1.6	0.1101	
RSCPU_HEK_SRR	-0.39351	0.40322	-0.976	0.3296	
RSCPU_HEK_ERX	0.73974	0.4682	1.58	0.1148	
tRNA	-4.70107	3.00585	-1.564	0.1185	
MFE_151	-0.02097	0.0105	-1.997	0.0464	*
Pair-prob	0.22075	0.16498	1.338	0.1815	
Lag3	0.05156	0.04475	1.152	0.2498	
Lag2	0.03957	0.04945	0.8	0.4239	
Lag1	0.47339	0.04486	10.554	<2e-16	***
<b>CO F9</b>					
(Intercept)	-0.428561	0.470664	-0.911	0.362978	
RSCU_Genome	0.987203	0.686076	1.439	0.150808	
RSCU_HEK_SRR	-0.510807	0.884671	-0.577	0.563933	
RSCU_HEK_ERX	-0.562106	1.060021	-0.53	0.596157	
RSCPU_Genome	-0.100063	0.230617	-0.434	0.664555	
RSCPU_HEK_SRR	0.012383	0.335691	0.037	0.970589	
RSCPU_HEK_ERX	0.119879	0.396397	0.302	0.762458	
tRNA	-3.330872	2.315622	-1.438	0.150943	
MFE_151	-0.007955	0.008367	-0.951	0.342211	
Pair-prob	-0.130687	0.200669	-0.651	0.515183	
Lag3	-0.02156	0.045746	-0.471	0.637641	
Lag2	0.177367	0.048099	3.688	0.000252	***
Lag1	0.390892	0.045033	8.68	<2e-16	***
<b>ACTB</b>					
(Intercept)	-1.404031	0.545822	-2.572	0.0105	*
RSCU_Genome	-0.310705	0.916587	-0.339	0.7348	
RSCU_HEK_SRR	-0.796966	1.242992	-0.641	0.5218	
RSCU_HEK_ERX	1.025136	1.514433	0.677	0.4989	
RSCPU_Genome	-0.100309	0.265983	-0.377	0.7063	
RSCPU_HEK_SRR	-0.075492	0.43977	-0.172	0.8638	
RSCPU_HEK_ERX	0.273948	0.478317	0.573	0.5672	
tRNA	1.554097	2.632387	0.59	0.5553	
MFE_151	-0.008856	0.007942	-1.115	0.2655	
Pair-prob	0.642673	0.26481	2.427	0.0157	*
Lag3	0.107027	0.052224	2.049	0.0411	*
Lag2	0.05047	0.057602	0.876	0.3815	
Lag1	0.442478	0.052295	8.461	6.64E-16	***
<b>GAPDH</b>					
(Intercept)	-0.149095	0.462448	-0.322	0.74736	
RSCU_Genome	1.409989	0.828001	1.703	0.08955	.
RSCU_HEK_SRR	-1.172997	1.265043	-0.927	0.3545	
RSCU_HEK_ERX	-0.093217	1.421276	-0.066	0.94775	
RSCPU_Genome	-0.652895	0.287641	-2.27	0.02388	*
RSCPU_HEK_SRR	-0.4377	0.438042	-0.999	0.31844	
RSCPU_HEK_ERX	1.113346	0.497467	2.238	0.0259	*
tRNA	-2.032467	2.679425	-0.759	0.44868	
MFE_151	0.008537	0.006816	1.252	0.2113	
Pair-prob	0.061488	0.22621	0.272	0.78594	
Lag3	0.031727	0.05551	0.572	0.56802	
Lag2	0.179942	0.058943	3.053	0.00246	**
Lag1	0.371884	0.056404	6.593	1.76E-10	***