RiboAI
-
Ribo-Seq
meets
Machine Learning

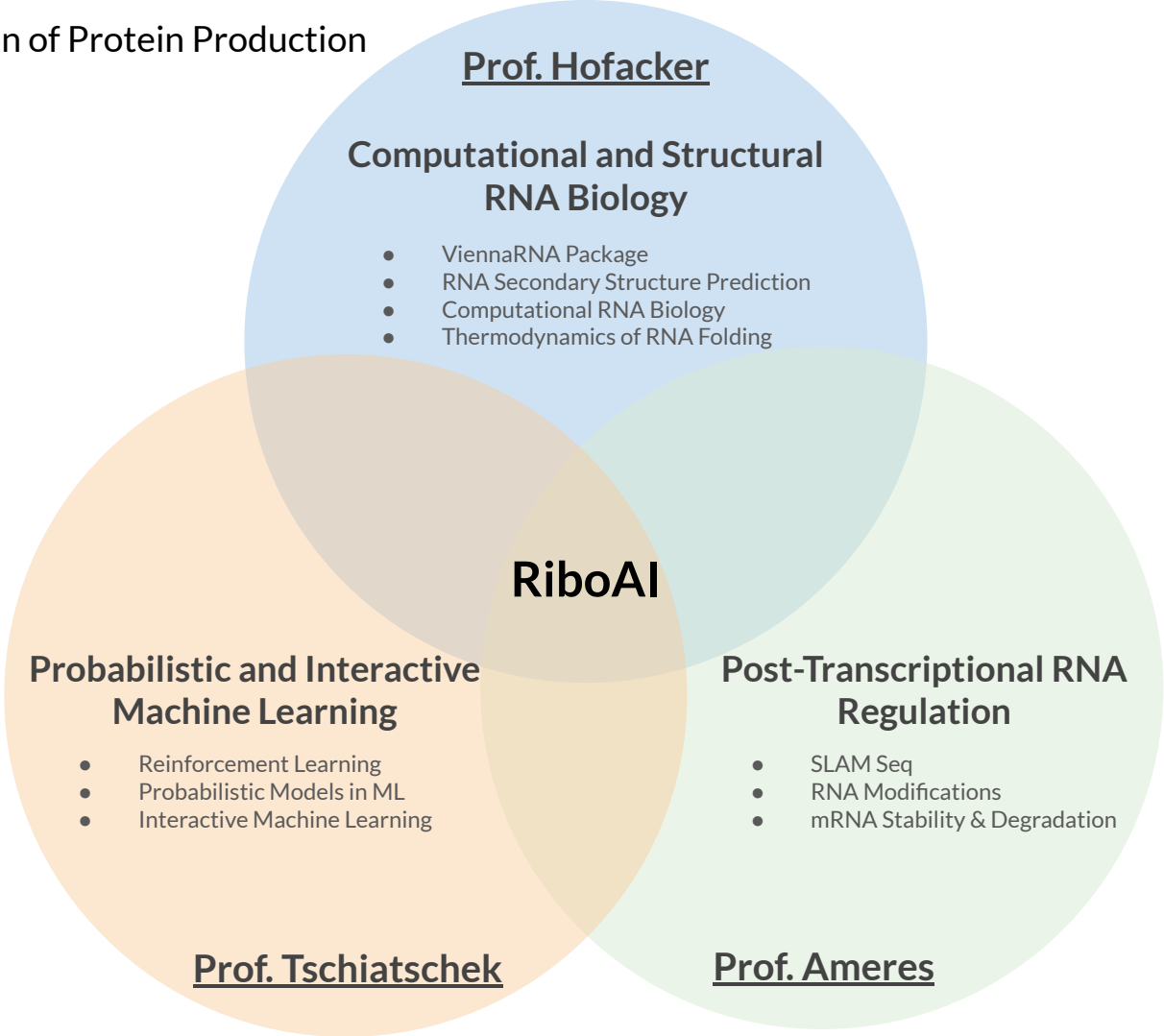
by:
Denis Skibinski

supervised by:
Prof. Hofacker

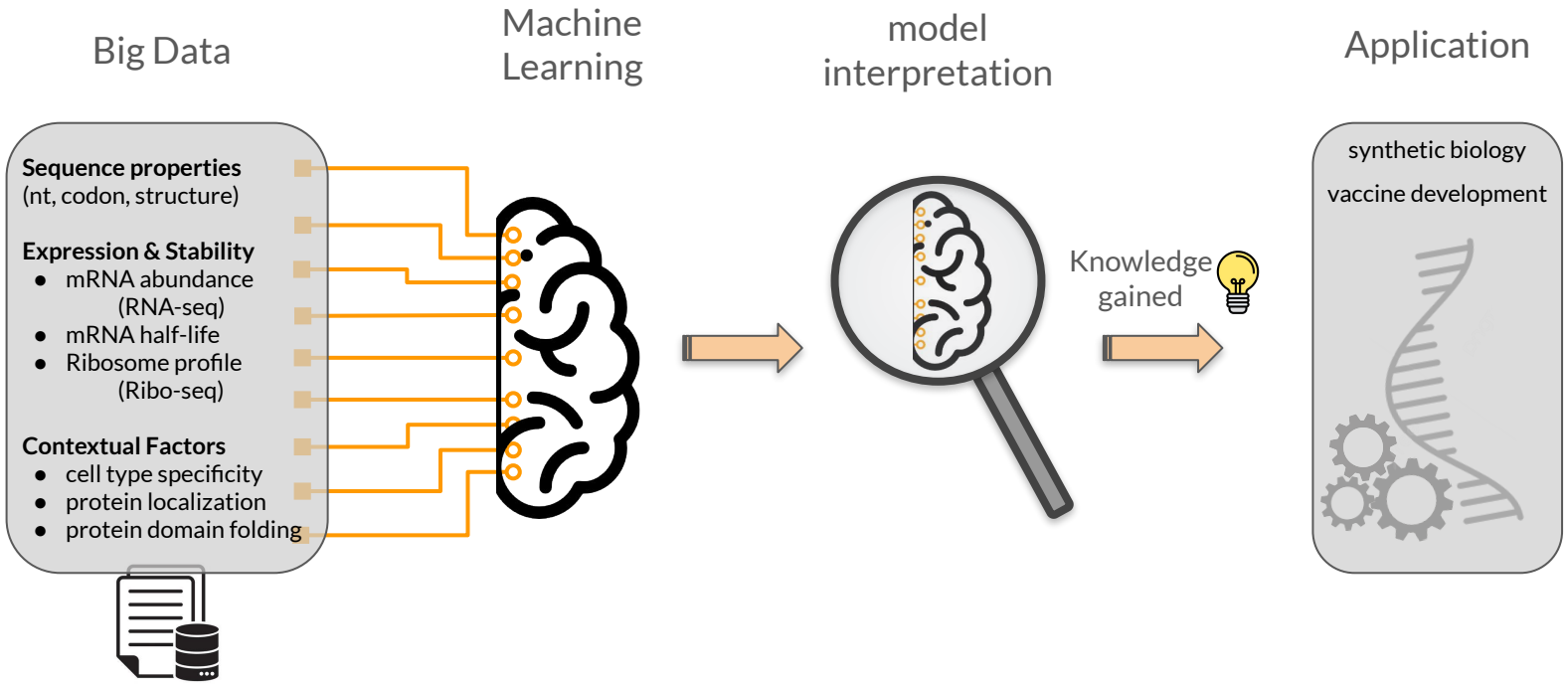
University Vienna
TBI

RiboAI: Data-Driven Exploration of Protein Production

3 Professors
+
1 Senior Scientist
+
2 PhD Student
=
1 RiboAI project



RiboAI: Data-Driven Exploration of Protein Production



RiboAI - PhD Timeline

START | 07/2024 – 1st Year | 07/2025 – 2nd Year | 07/2026 – 3rd Year | 07/2027 – 07/2028 | END

today
02/2025

Data generation

- Sourcing/Generation of Data
- QC of Data
- Implementation of baselines

Machine Learning

- Developing advanced ML models (Gabriele Martino, 2nd PhD student)
- Training these models

Explainable AI

- Adoption of XAI techniques
- Theoretical verification of obtained biological knowledge

Validation

- Experimental validation (in-vitro)
- Model refinement

Data (Hot-)encoded, the machine vision

per transcript: Nucleotide-Resolution Features

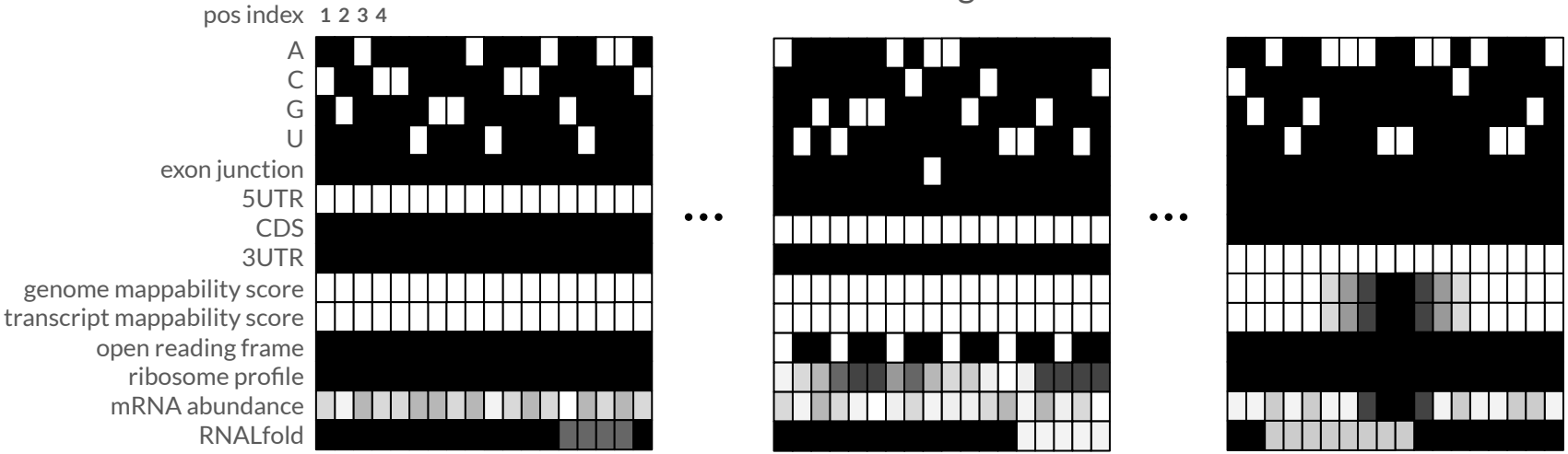
position	A	C	G	U	exon junciton	5UTR	RNALfold	mRNA abundance	ribosome profile
1	0	1	0	0	0	1	-5.6	0.3	0
2	0	0	1	0	0	1	-5.6	0.8	0
3	1	0	0	0	0	1	-5.6	0.2	0
4	0	0	1	0	0	1	-5.6	1.8	0

target

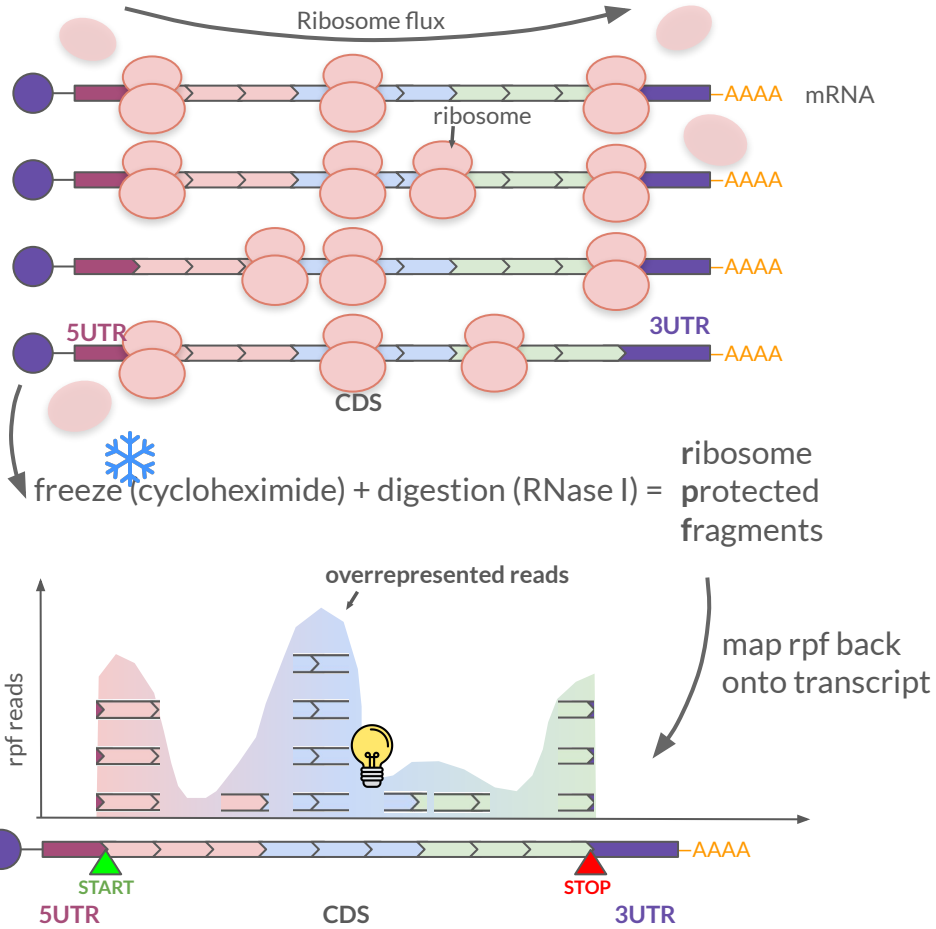
...

⋮

Embedding:



Ribosome Profile - snapshot of translation in action

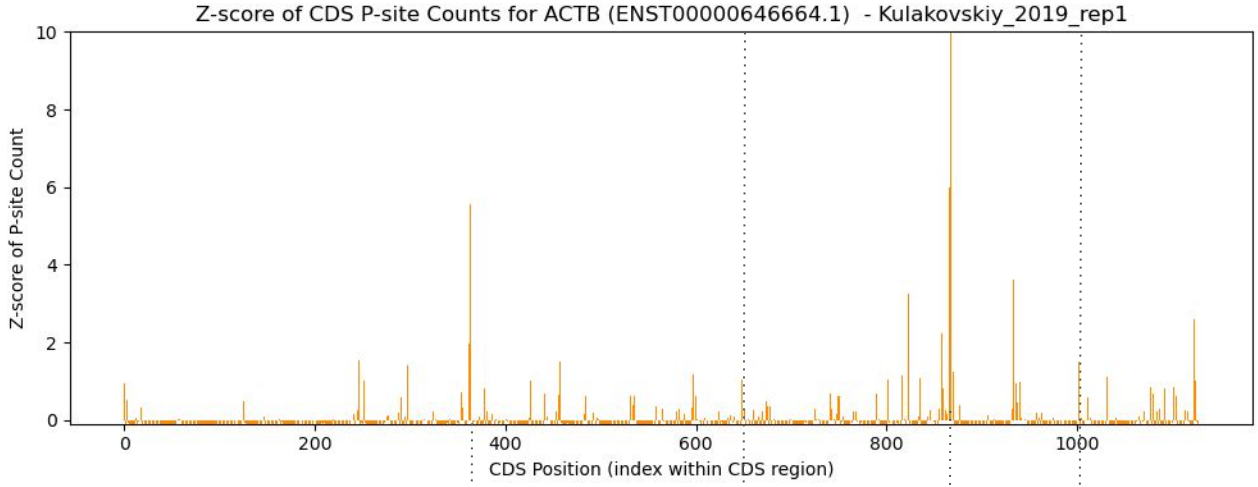


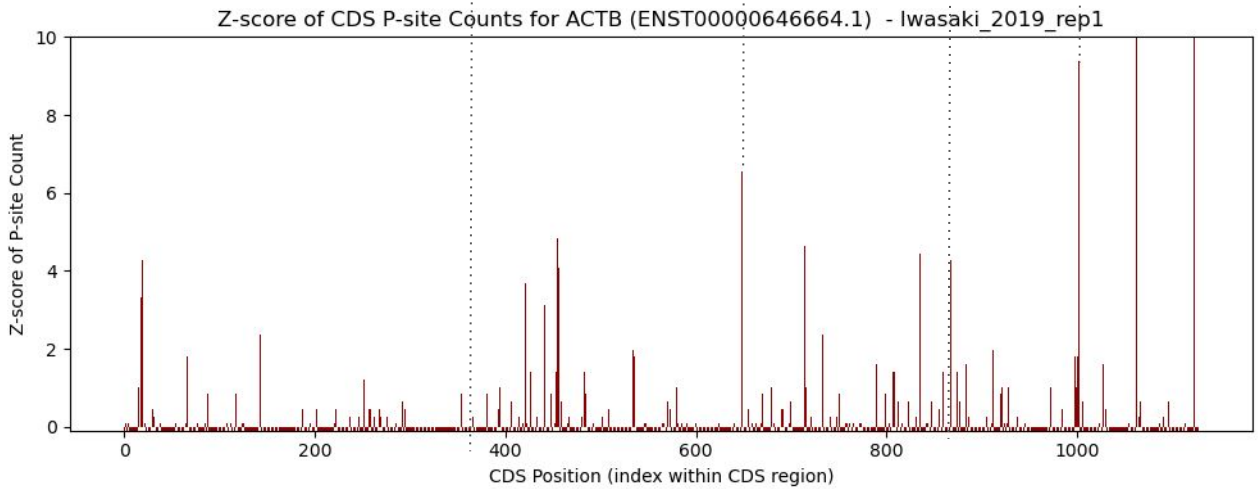
How-to ribosome profile:

1. created by freezing/stopping ribosomes on mRNA
2. digestion of not protected mRNA around ribosome-> creating ribosome protected fragments (reads)
3. sequence, filter & map reads back onto the transcript

Ingolia NT, Hussmann JA, Weissman JS.
Ribosome Profiling: Global Views of Translation.
Cold Spring Harb Perspect Biol. 2019 May 1;11(5):a032698.
doi: 10.1101/cshperspect.a032698.

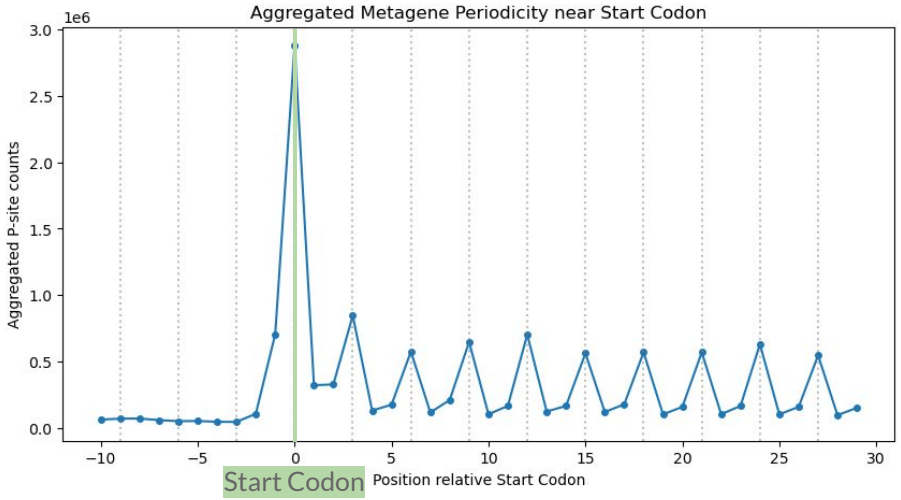
Ribo-seq **low** reproducibility - between datasets





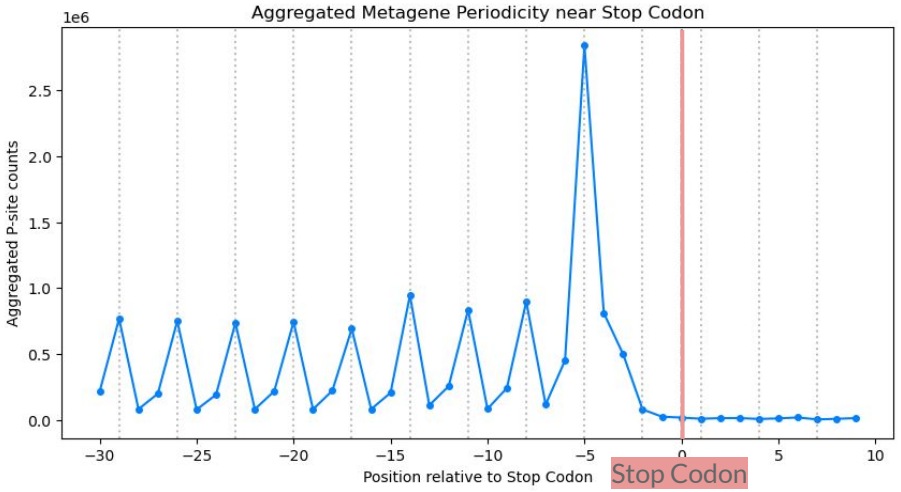
- all HEK datasets
- Low peak reproducibility around (p=.2) between datasets
- need to assure high quality ribosome profiles

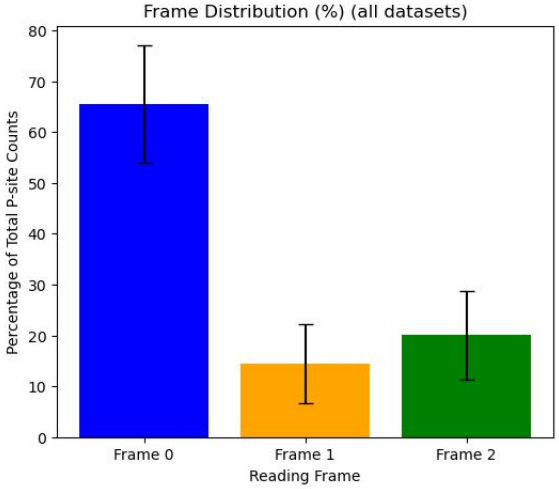
Quality Control - Ribo-seq



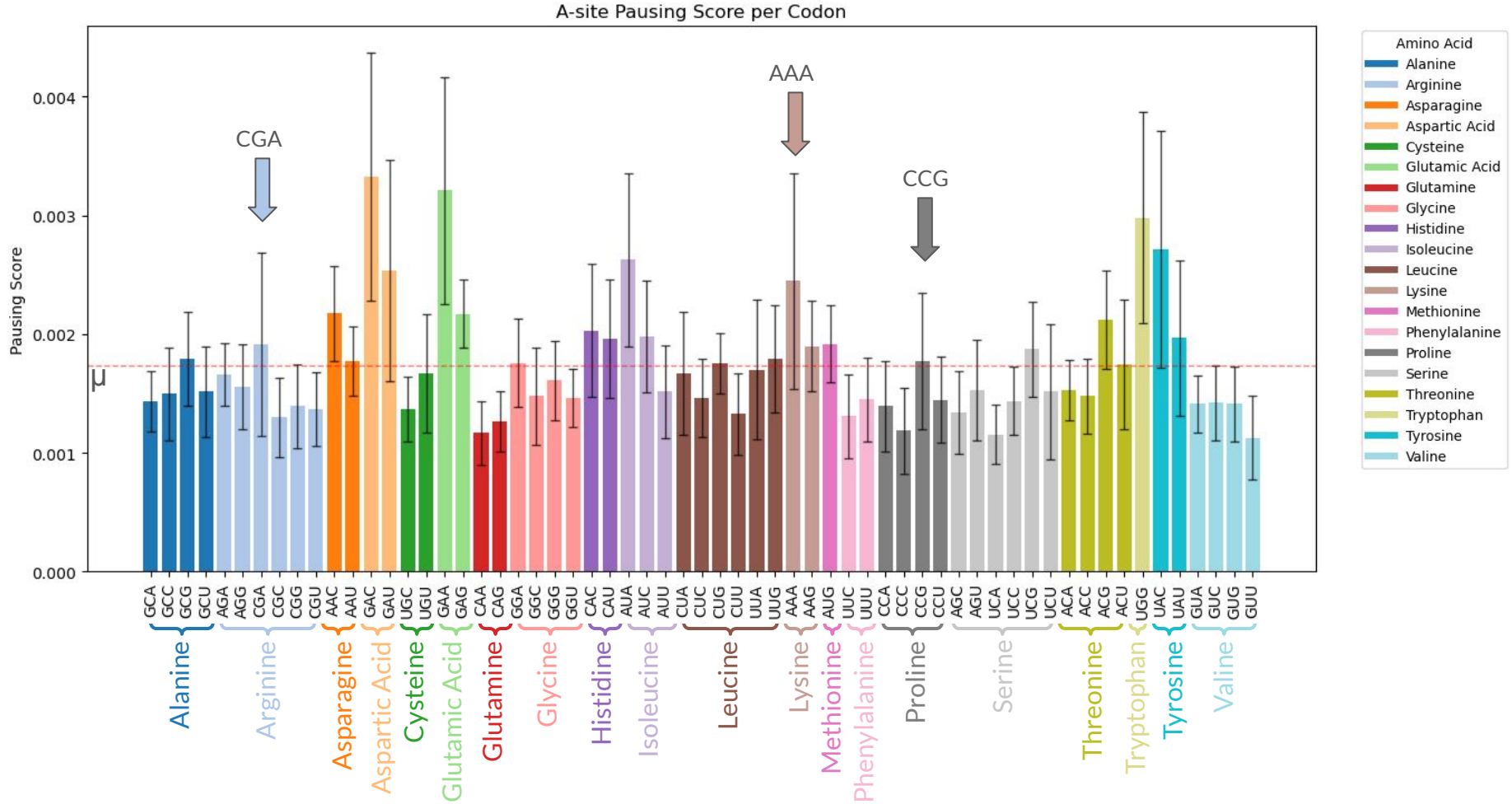
Indicator of high quality Ribosome Profile Data:

- no mapping outside of CDS
- clean cut at Start Codon/ before Stop Codon
- mean % of reads mapped in OpenReadingFrame (frame 0) > 60%
- high freq every 3nt in CDS

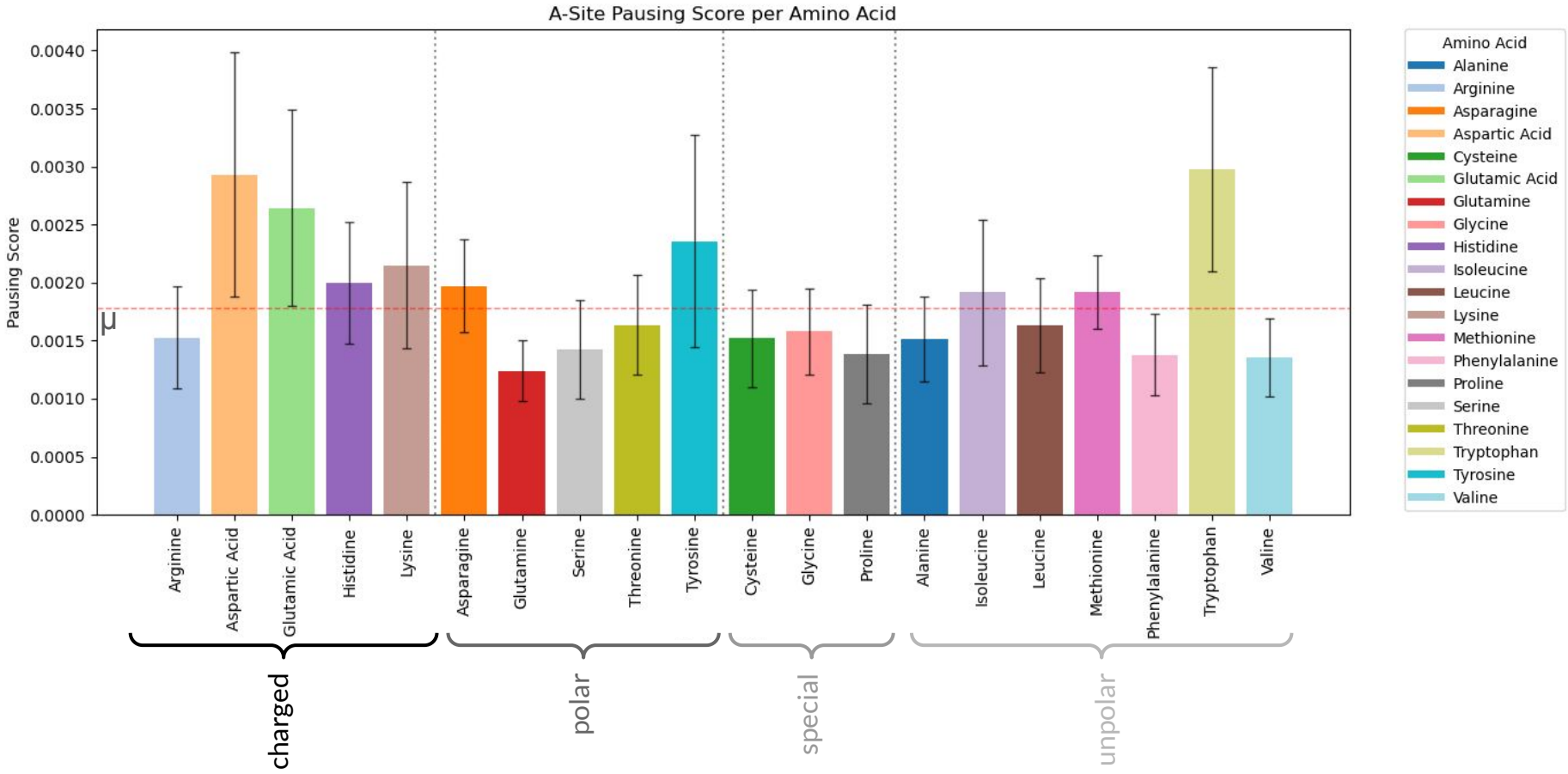




QC - Pausing Score - A-site Codon - sanity check



QC - Pausing Score - Amino Acid - sanity check



Translation Efficiency in Feature space

per transcript:

Nucleotide-Level-Resolution Features

position	A	C	G	U	exon border	5UTR	RNALfold	mRNA abundance	ribosome profile
1	0	1	0	0	0	1	-5.6	0.3	0
2	0	0	1	0	0	1	-5.6	0.8	0
3	1	0	0	0	0	1	-5.6	0.2	0
4	0	0	1	0	0	1	-5.6	1.8	0

calculated out of Ribo-seq

targets

Transcript-Level Features

Tx_ID	Ribo_RPKM	RNA_RPKM	GC_Content	A:AU_Ratio	C:CG_Ratio	Exon Junct Density	HalfLife	Translation Efficiency	Protein abundance
ENST00001	16.5	74.3	0.43	0.95	0.87	2	3.4	1.4	1.4
ENST00002	2.8	12.1	0.52	1.11	0.90	4	1.5	0.8	0.8

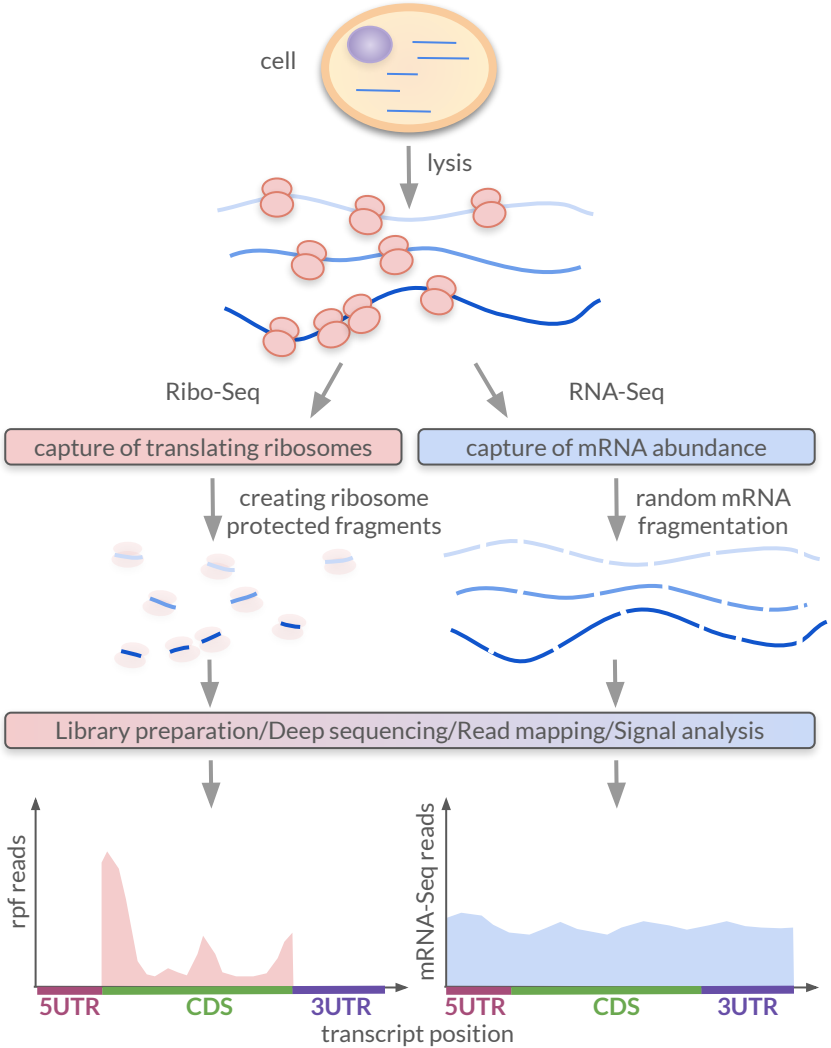
Translation Efficiency

Calculating Translation Efficiency:

$$log(TE_{tx}) = log\left(\frac{RiboSeq_{tx}}{RNASeq_{tx}}\right)$$

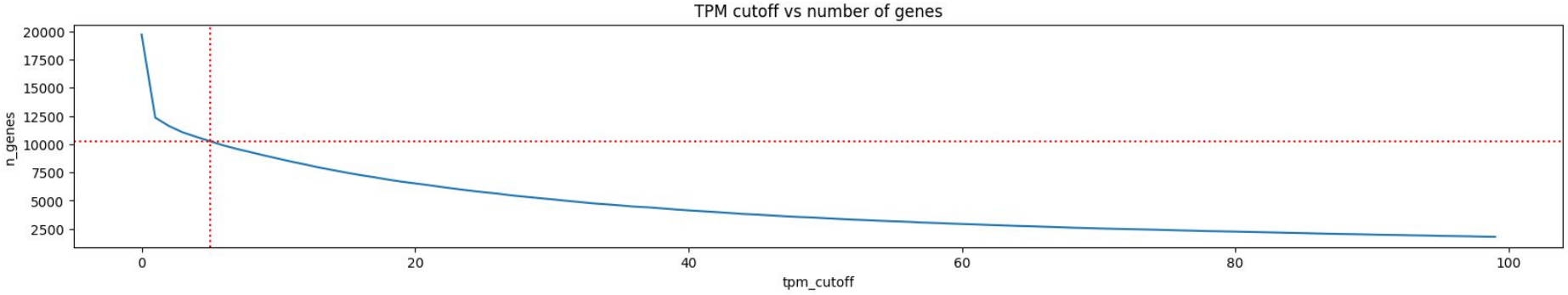
High TE: Suggests strong protein synthesis

Low TE: Suggest weak protein synthesis

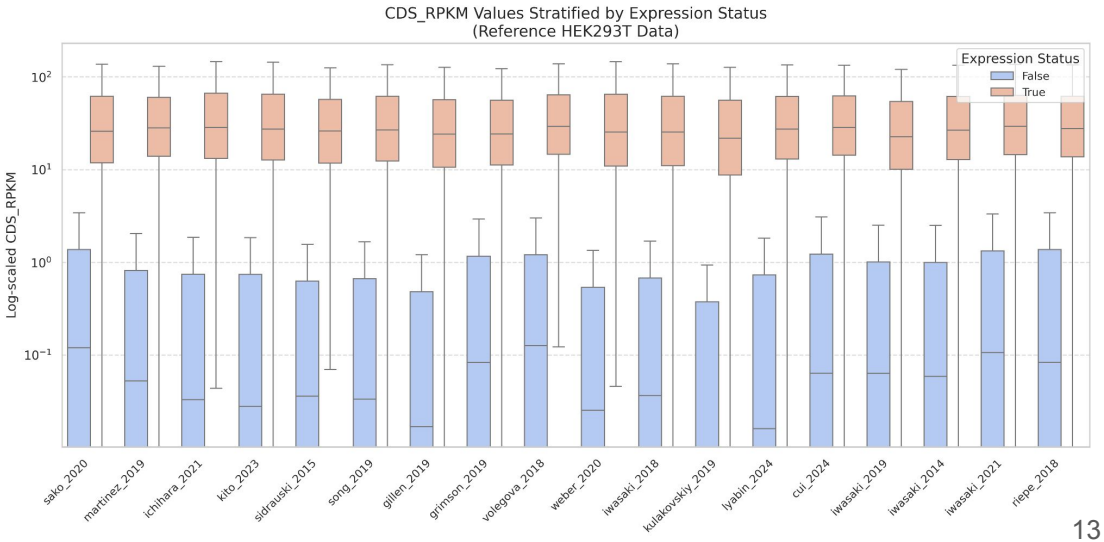


Brar, G., Weissman, J.
Ribosome profiling reveals the what, when, where and how of protein synthesis.
Nat Rev Mol Cell Biol 16, 651–664 (2015)
https://doi.org/10.1038/nrm4069

Improve accuracy through highly expressed Genes in HEK cells



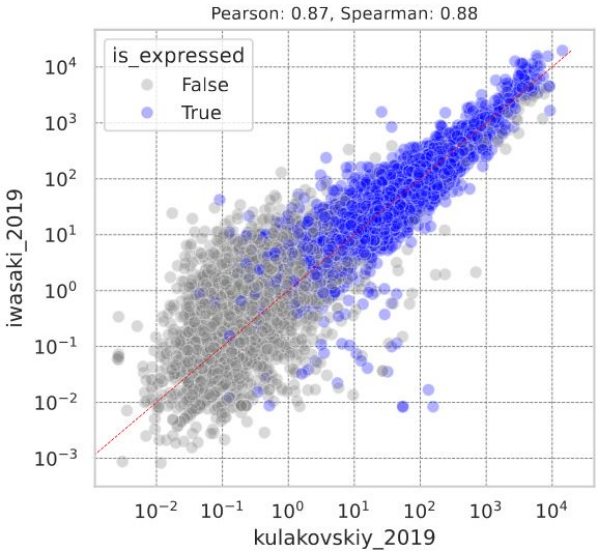
tpm_cutoff = 5
results in 10255 highly expressed genes



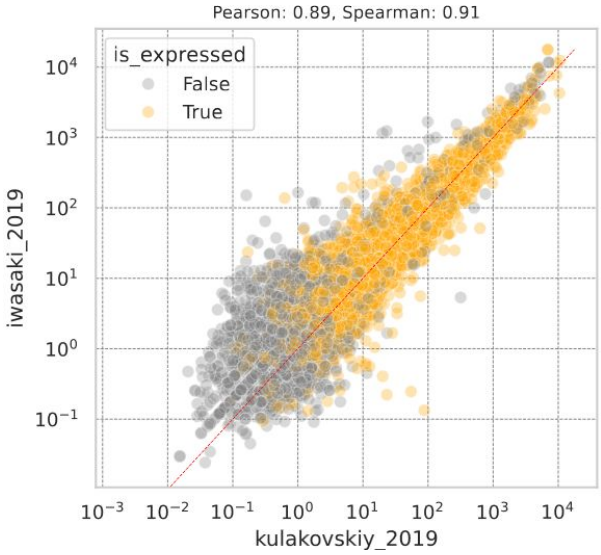
PE75 RNAseq data of WT HEK293T cells (SAMN22420269-SAMN22420271)
GEO accession: GSE186192

Cut-off leads to higher correlation between datasets (all HEK)

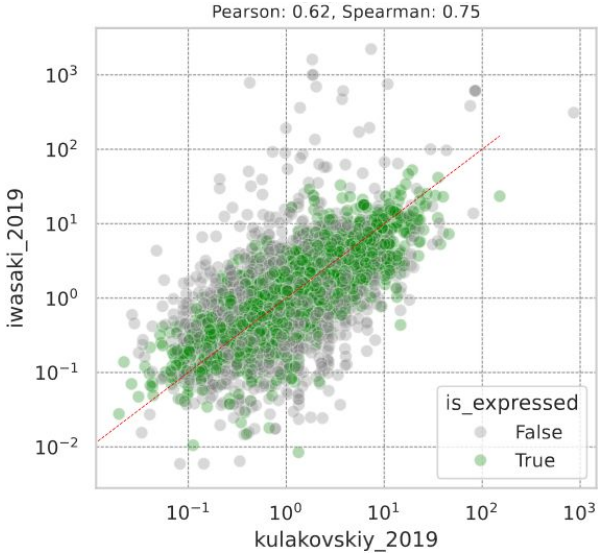
mRNA abundance (rpkm)



ribosome occupancy (rpkm)



translation efficiency



Correlation - Translation Efficiency

TE correlation (Spearman)
Median: 0.60

	song_2019	sako_2020	martinez_2019	sidrauski_2015	ichihara_2021	weber_2020	kito_2023	volegova_2018	cui_2024	lyabin_2024	riepe_2018	grimson_2019	gillen_2019	kulakovskiy_2019	iwasaki_2019	iwasaki_2018	iwasaki_2021	iwasaki_2014
song_2019	1	0.65	0.69	0.67	0.59	0.47	0.49	0.57	0.52	0.58	0.42	0.32	0.31	0.55	0.56	0.57	0.73	0.66
sako_2020	0.65	1	0.56	0.48	0.52	0.53	0.46	0.44	0.46	0.44	0.52	0.26	0.2	0.52	0.53	0.54	0.57	0.57
martinez_2019	0.69	0.56	1	0.63	0.43	0.28	0.3	0.37	0.51	0.51	0.38	0.47	0.34	0.68	0.57	0.65	0.74	0.68
sidrauski_2015	0.67	0.48	0.63	1	0.48	0.38	0.44	0.55	0.49	0.59	0.44	0.49	0.45	0.58	0.62	0.6	0.72	0.69
ichihara_2021	0.59	0.52	0.43	0.48	1	0.7	0.81	0.78	0.61	0.74	0.54	0.44	0.52	0.41	0.56	0.53	0.64	0.66
weber_2020	0.47	0.53	0.28	0.38	0.7	1	0.84	0.76	0.62	0.7	0.63	0.54	0.62	0.46	0.59	0.53	0.52	0.63
kito_2023	0.49	0.46	0.3	0.44	0.81	0.84	1	0.87	0.66	0.81	0.64	0.55	0.66	0.41	0.58	0.54	0.6	0.66
volegova_2018	0.57	0.44	0.37	0.55	0.78	0.76	0.87	1	0.71	0.86	0.58	0.58	0.7	0.45	0.62	0.53	0.67	0.71
cui_2024	0.52	0.46	0.51	0.49	0.61	0.62	0.66	0.71	1	0.78	0.54	0.63	0.65	0.6	0.6	0.55	0.65	0.7
lyabin_2024	0.58	0.44	0.51	0.59	0.74	0.7	0.81	0.86	0.78	1	0.57	0.63	0.75	0.55	0.64	0.61	0.74	0.76
riepe_2018	0.42	0.52	0.38	0.44	0.54	0.63	0.64	0.58	0.54	0.57	1	0.51	0.52	0.6	0.64	0.63	0.6	0.65
grimson_2019	0.32	0.26	0.47	0.49	0.44	0.54	0.55	0.58	0.63	0.63	0.51	1	0.8	0.73	0.73	0.64	0.59	0.73
gillen_2019	0.31	0.2	0.34	0.45	0.52	0.62	0.66	0.7	0.65	0.75	0.52	0.8	1	0.55	0.65	0.6	0.6	0.69
kulakovskiy_2019	0.55	0.52	0.68	0.58	0.41	0.46	0.41	0.45	0.6	0.55	0.6	0.73	0.55	1	0.75	0.72	0.69	0.76
iwasaki_2019	0.56	0.53	0.57	0.62	0.56	0.59	0.58	0.62	0.6	0.64	0.64	0.73	0.65	0.75	1	0.86	0.76	0.86
iwasaki_2018	0.57	0.54	0.65	0.6	0.53	0.53	0.54	0.53	0.55	0.61	0.63	0.64	0.6	0.72	0.86	1	0.76	0.8
iwasaki_2021	0.73	0.57	0.74	0.72	0.64	0.52	0.6	0.67	0.65	0.74	0.6	0.59	0.6	0.69	0.76	0.76	1	0.86
iwasaki_2014	0.66	0.57	0.68	0.69	0.66	0.63	0.66	0.71	0.7	0.76	0.65	0.73	0.69	0.76	0.86	0.8	0.86	1

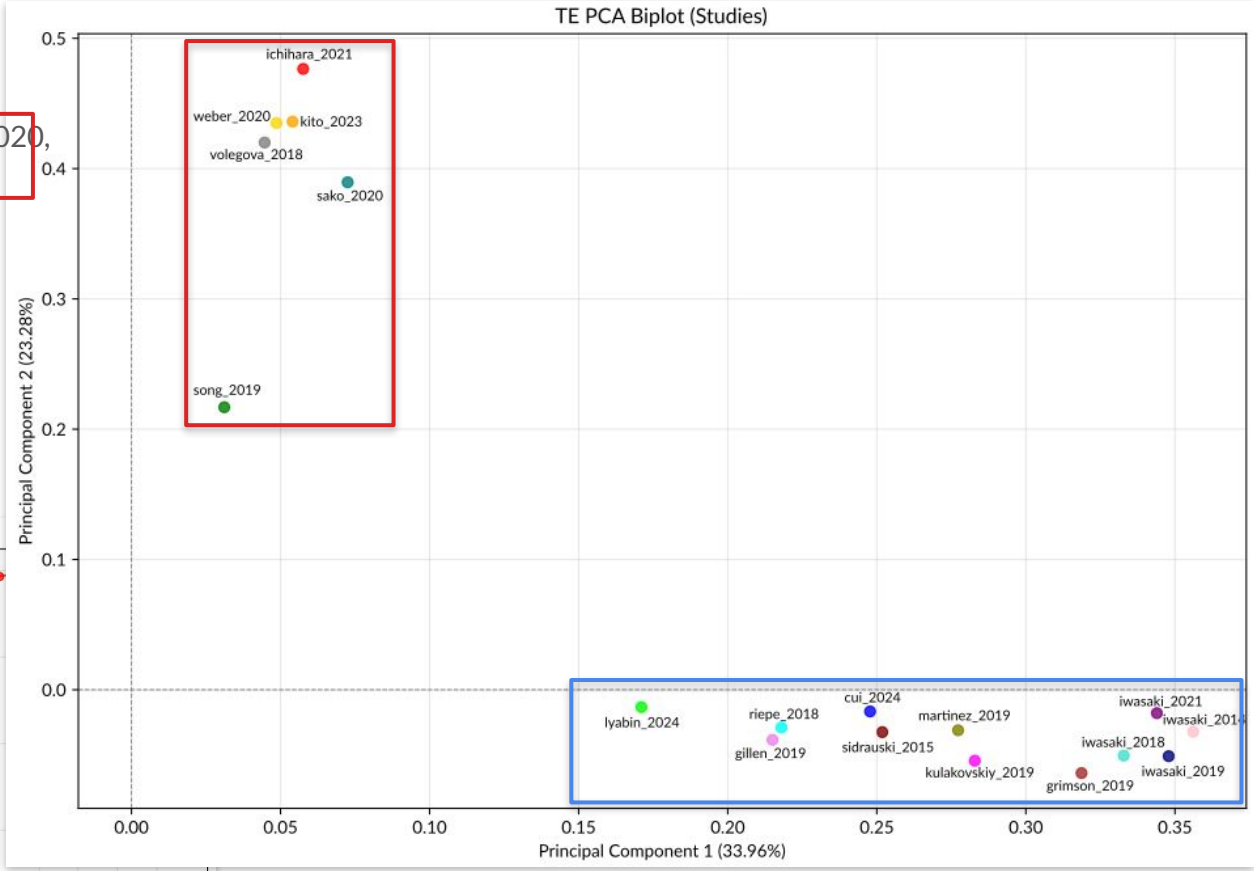
1.0
0.8
0.6
0.4
0.2
0.0

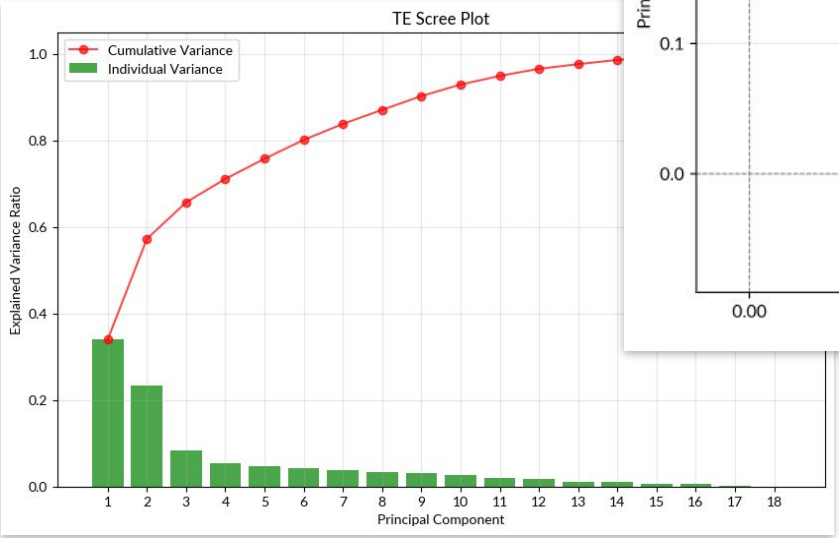
unexpected clustering

expected clustering

Looking deeper: PCA - TE

ichihara_2021, kito_2023, volegova_2018, sako_2020, weber_2020, song_2019 - PC2 cluster

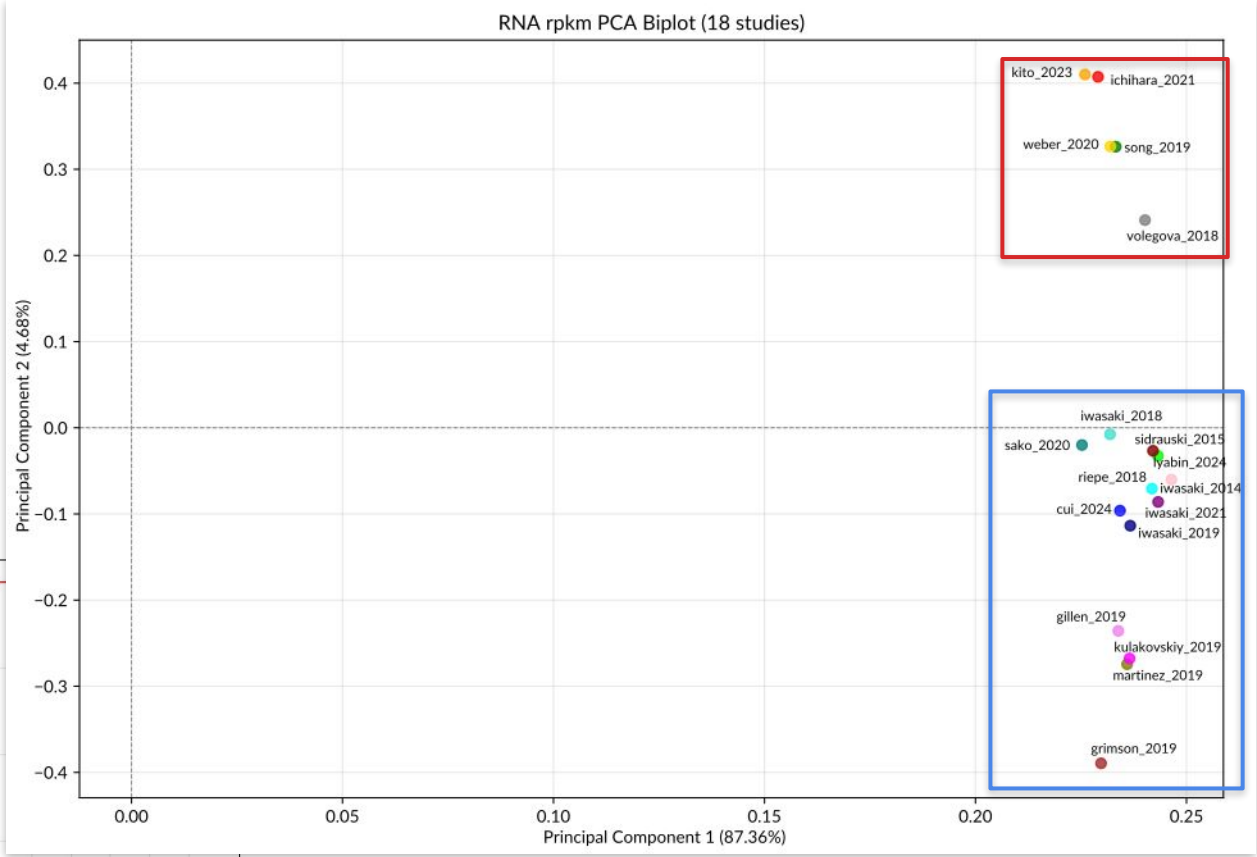


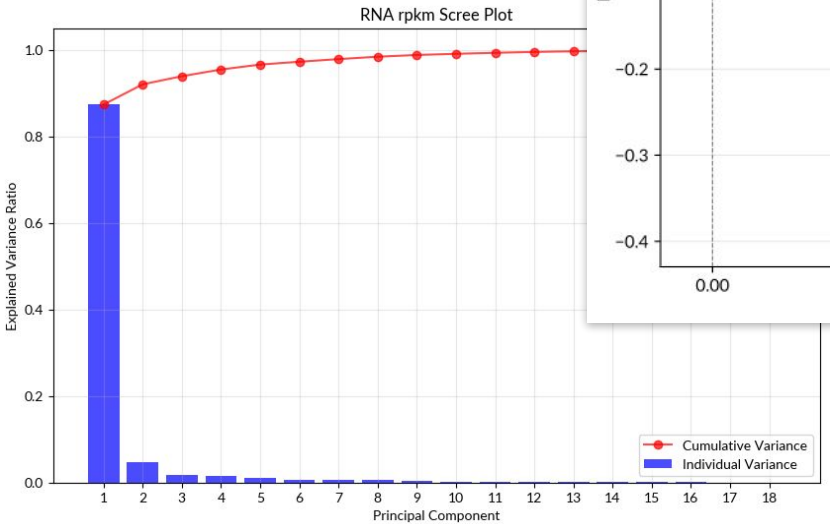


rest - PC1 cluster

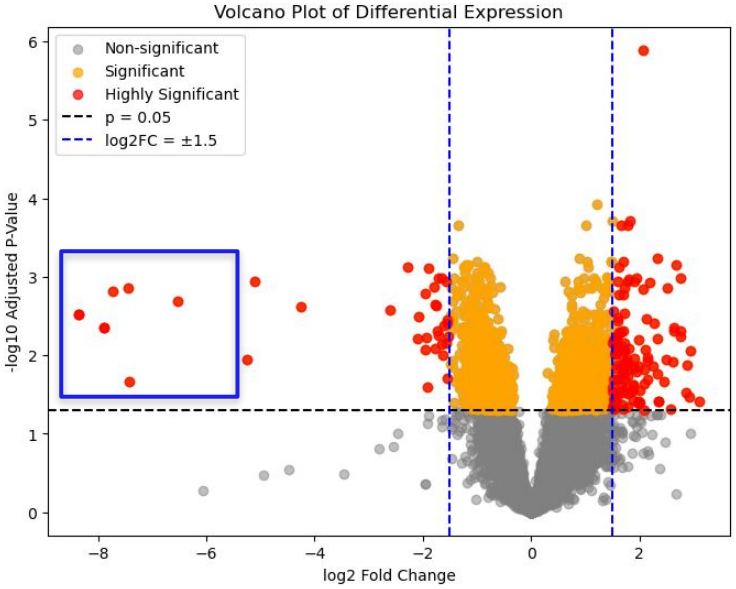
PCA - mRNA abundance (rpkm)

Same clustering as in Translation Efficiency PCA

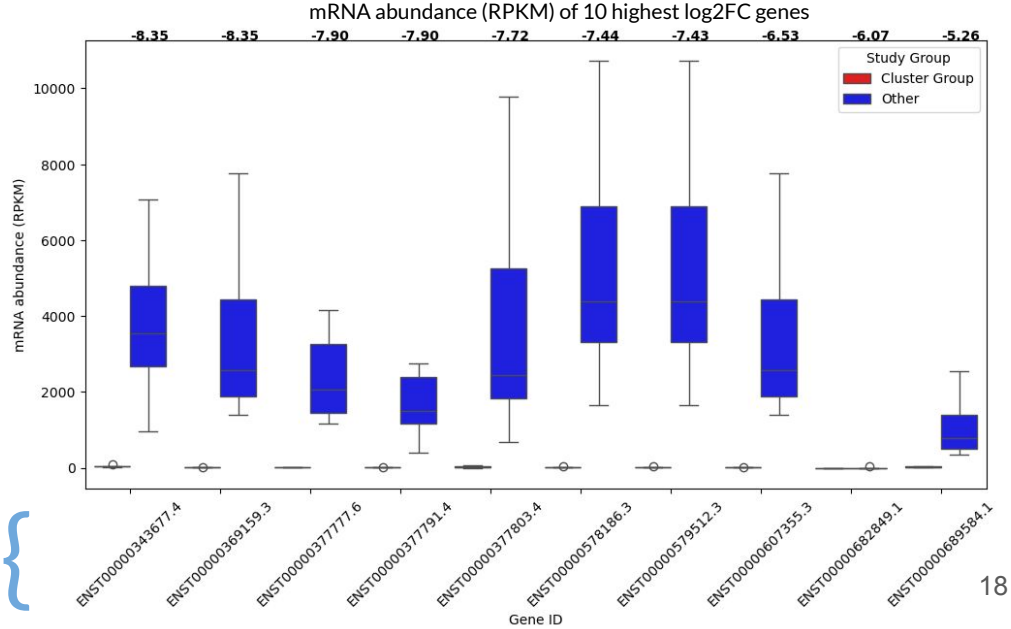




PCA - differential expression analysis - do a few genes drive the PC2?

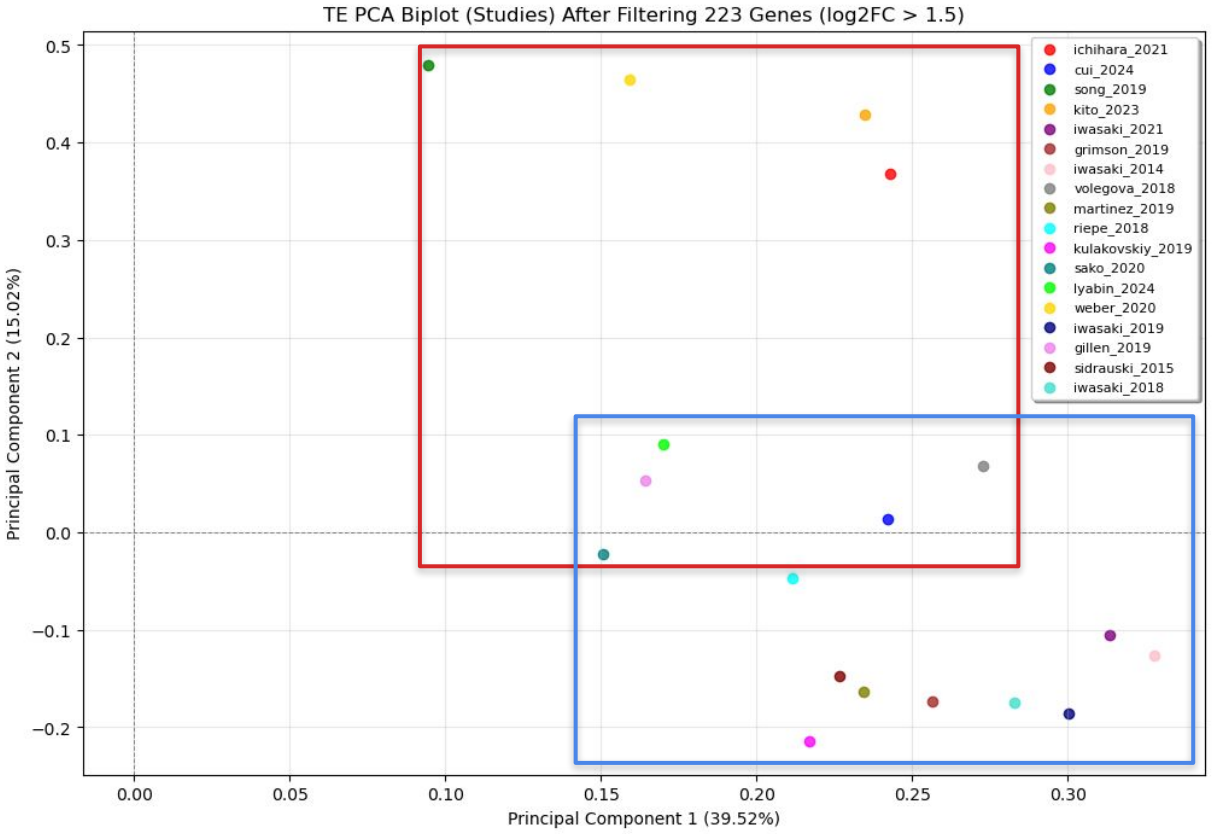


ʔ I could find no explanation in papers ʔ



histone-encoding genes
(H1, H2, H3, H4 variants)

PCA - removing the culprits



- PC2 cluster dissolves
- PC2 decreased/PC1 increased

Maybe the machine does not care - time will tell!

per transcript:

Nucleotide-Resolution-Level Features

position	A	C	G	U	exon border	5UTR	RNALfold	mRNA abundance	ribosome profile
1	0	1	0	0	0	1	-5.6	0.3	0
2	0	0	1	0	0	1	-5.6	0.8	0
3	1	0	0	0	0	1	-5.6	0.2	0
4	0	0	1	0	0	1	-5.6	1.8	0

targets

Transcript-Level Features

Tx_ID	Ribo_RPKM	RNA_RPKM	GC_Content	A:AU_Ratio	C:CG_Ratio	Exon Junct Density	HalfLife	Translation Efficiency	Protein abundance
ENST00001	16.5	74.3	0.43	0.95	0.87	2	3.4	1.4	1.4
ENST00002	2.8	12.1	0.52	1.11	0.90	4	1.5	0.8	0.8

Going forward…

Ribo-Seq Data QC is challenging

Variability across datasets requires careful Data curation

which features to add/discard for ML

which genes/transcripts to keep for ML

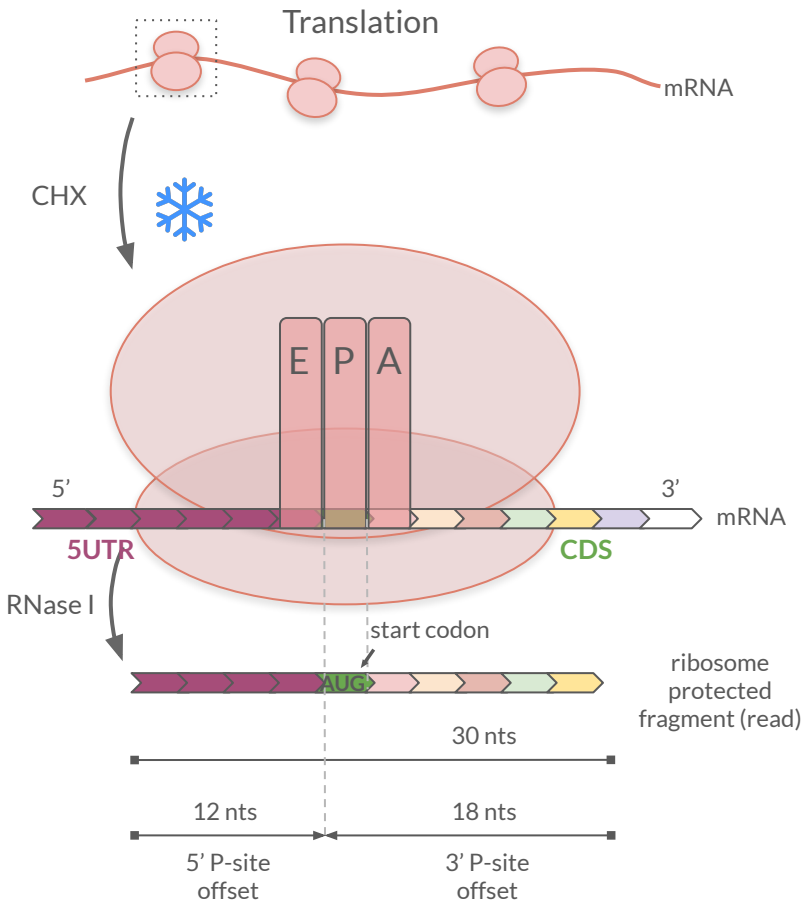
what ML architecture to choose? (nucleotide level vs transcript level)

Thanks to…

Ivo Hofacker (supervisor)
Gabriele Martino (PhD in crime)
Niko Popitsch (Senior Prophet Scientist at AmeresLab)
TBI (emotional support group)

And you for listening 

Problem with variability in Ribo-seq read length



Accurate P-site localization is the foundation of interpreting ribosome profiling data

Variability in rpf lengths! caused by: experimental conditions, ribosome conformations and nuclease biases

→ riboWaltz: most accurate p-site finder!

Aligns reads of the same length and finds with the help of the start codon the ORF and the P-site offsets, which are generalized over the whole rpf length bin.

Lauria F, Tebaldi T, Bernabò P, Groen EJN, Gillingwater TH, Viero G (2018) riboWaltz: Optimization of ribosome P-site positioning in ribosome profiling data. PLoS Comput Biol 14(8): e1006169. https://doi.org/10.1371/journal.pcbi.1006169