

Scikit-ribo reveals precise codon-level translational control by dissecting ribosome pausing and codon elongation

Han Fang

October 26, 2016
Biological Data Science



The hidden treasure in genomics



The hidden treasure in genomics



Ribosome-Mediated Specificity

in Ho
Verte

Ribosome Profiling Reveals a Cell-Type-Specific

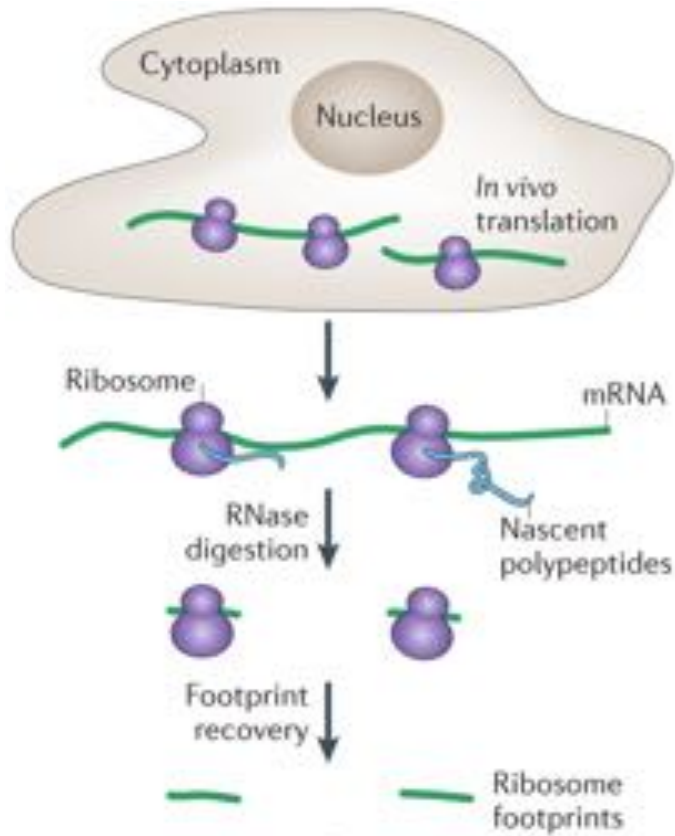
Tr

Dynamics of ribosome scanning and recycling

rev

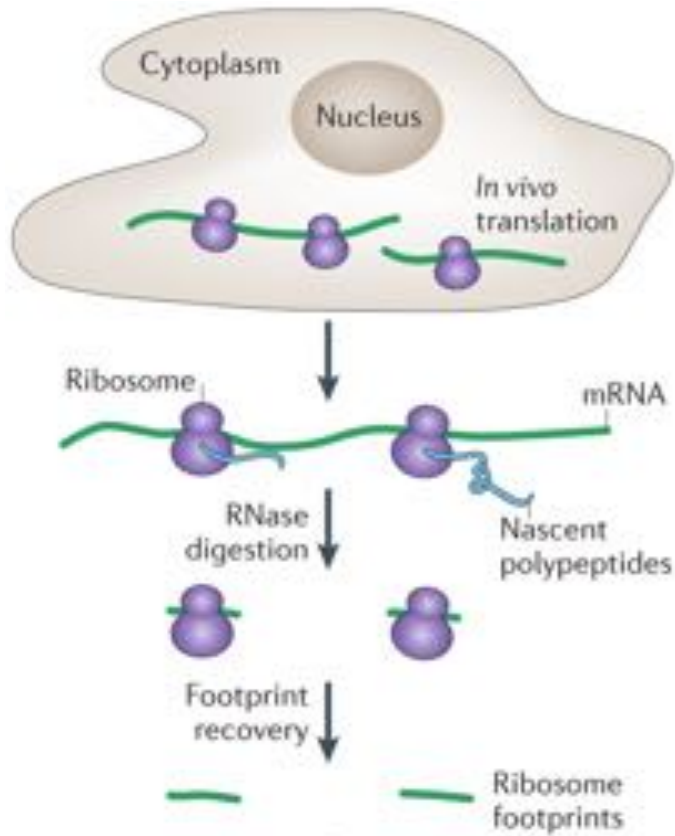
Ribosome profiling reveals features of normal and disease-associated mitochondrial translation

What is ribosome profiling (Riboseq)?

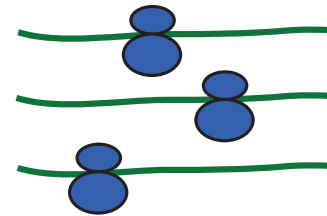


Ingolia et al. *Science*. (2009)
Ingolia. *Nat Rev Genet*. (2014)

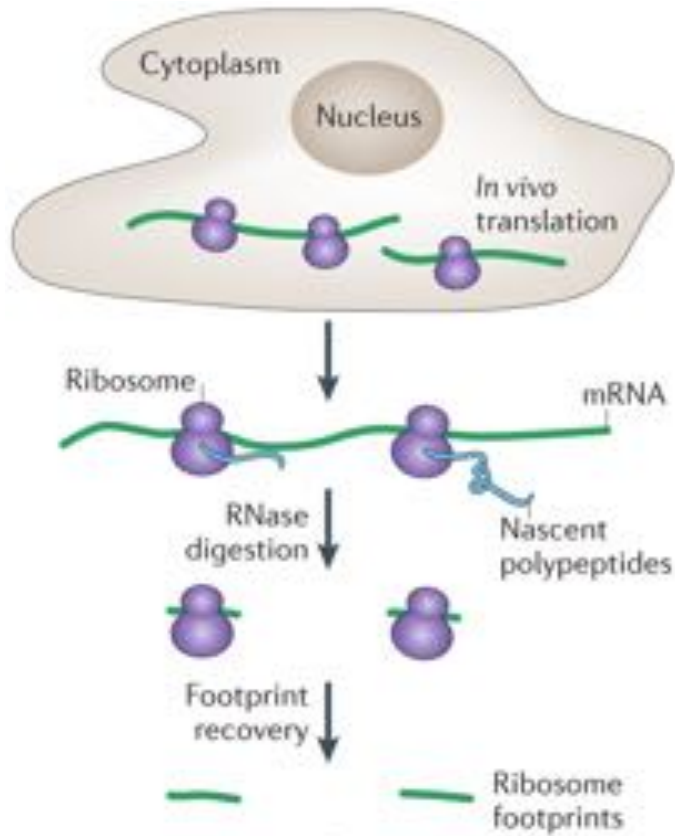
What is ribosome profiling (Riboseq)?



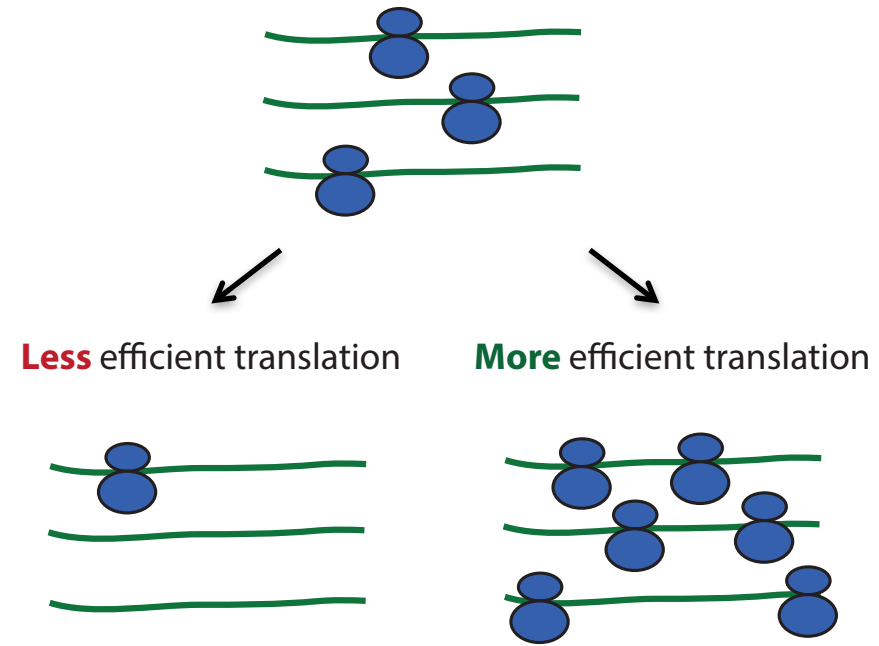
Normal translation efficiency (TE)



What is ribosome profiling (Riboseq)?

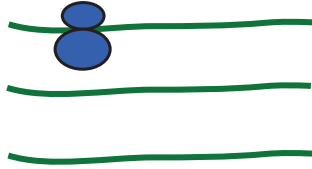


Normal translation efficiency (TE)



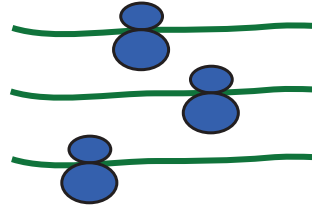
Calculate translational efficiency (TE)

Less efficient translation



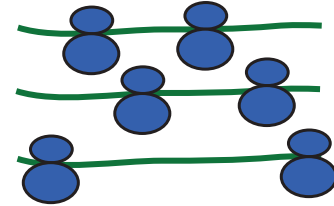
$$\log_2(TE) < 0$$

Normal translation efficiency (TE)



$$\log_2(TE) = 0$$

More efficient translation

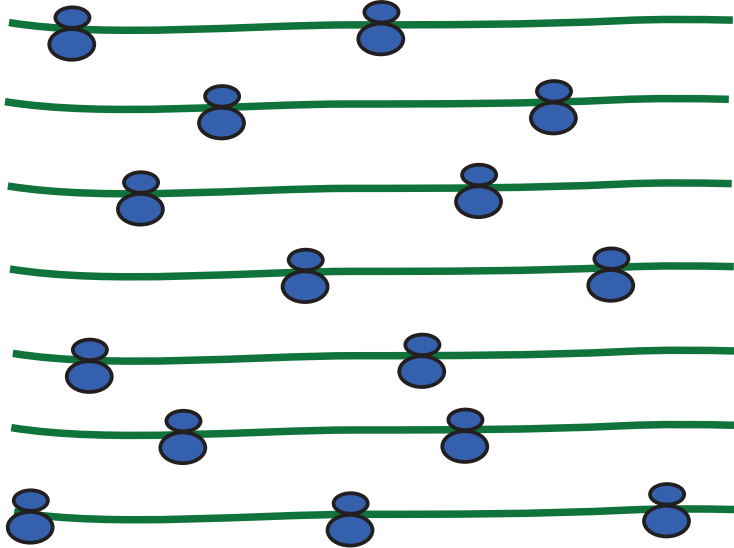


$$\log_2(TE) > 0$$

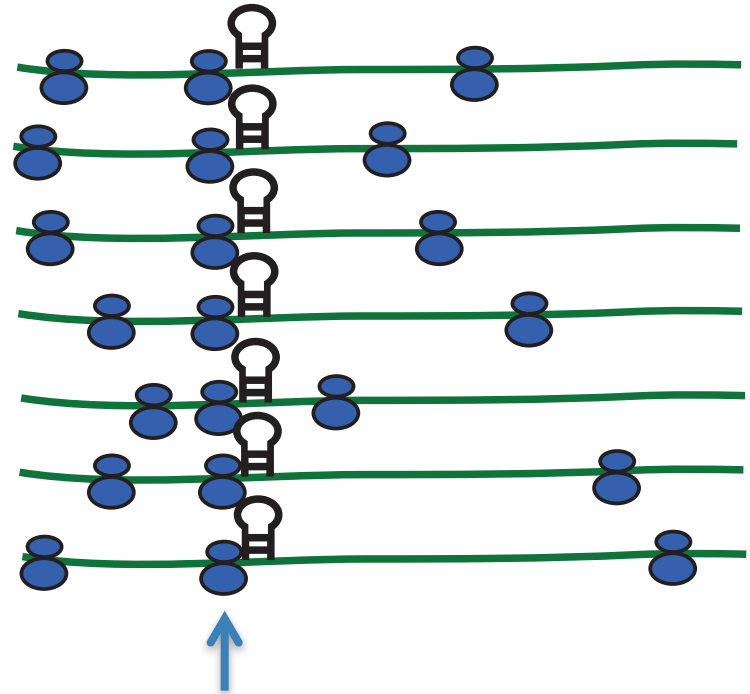
$$TE = \frac{\text{Riboseq rpkm}}{\text{RNAseq rpkm}}$$

Hypothesis: TE distribution could be skewed by ribosome pausing events.

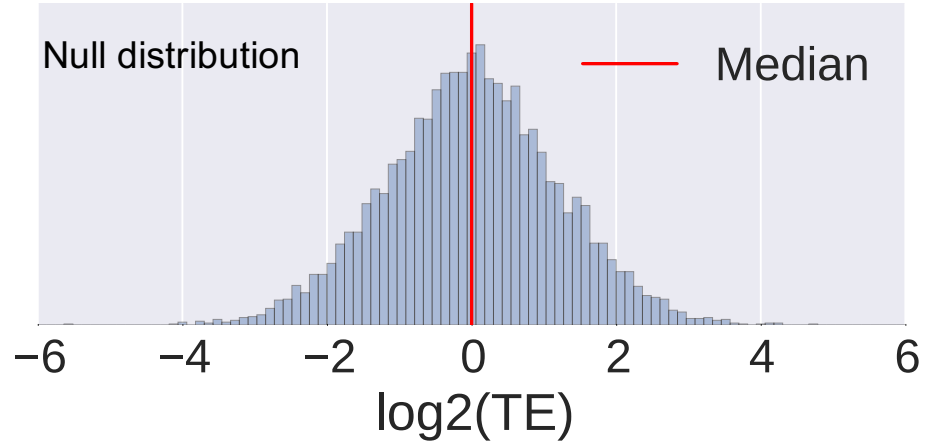
Ribosome footprints without bias



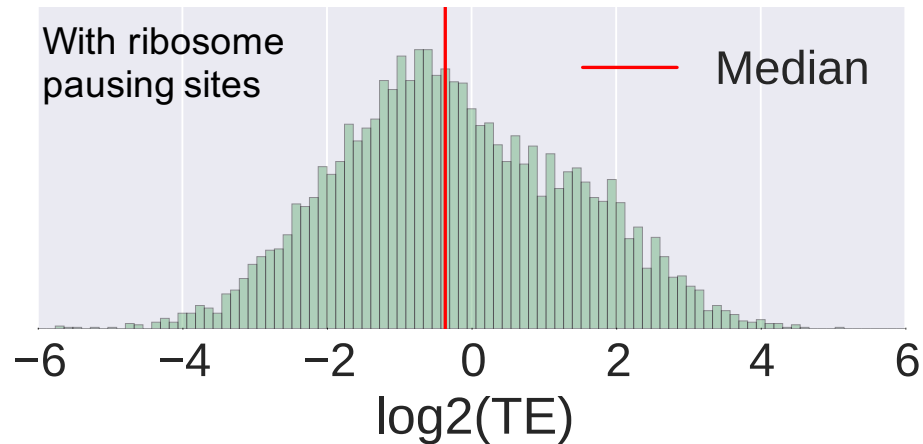
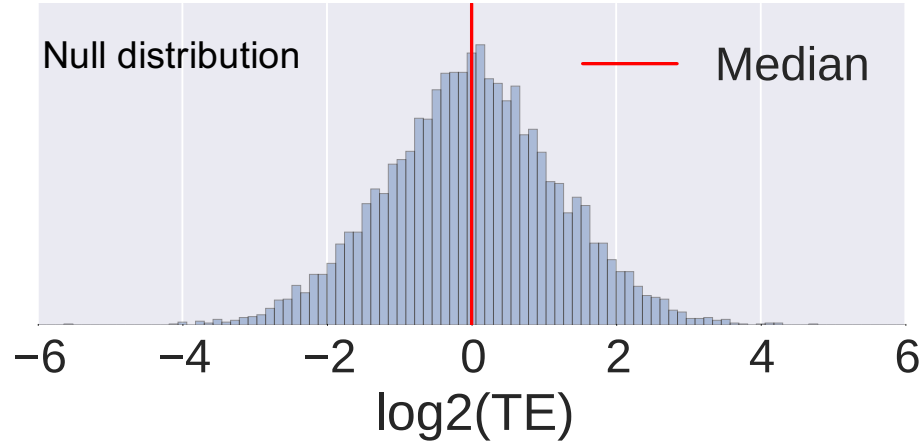
Ribosome footprints with pausing



Simulated *S. cerevisiae* data

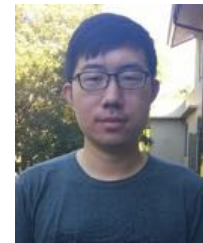


Simulated *S. cerevisiae* data - TE distribution are negatively-skewed by ribosome pausing events



Randomly imputed
ribosome pausing sites
to 20% of the genes

Ribosome pausing sites (peaks) finding by negative binomial mixture model



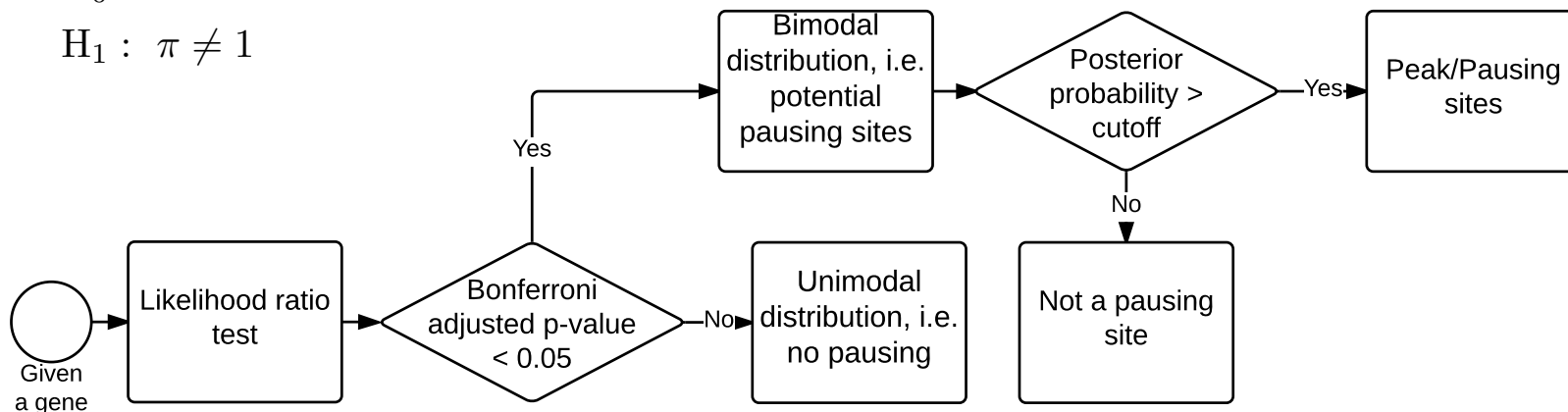
Yifei Huang

$$P(\mathbf{X}_i | \pi_i, \mu_i, k_i, r_i) = \prod_j \pi_i \mathcal{NB}(X_{ij} | \mu_i, r_i) + (1 - \pi_i) \mathcal{NB}(X_{ij} | k_i \mu_i, r_i),$$

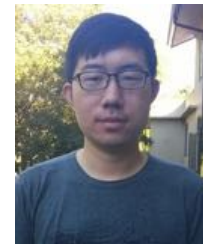
for gene i at position j , where $k \geq 5$

$H_0 : \pi = 1$

$H_1 : \pi \neq 1$



Ribosome pausing sites (peaks) finding by negative binomial mixture model



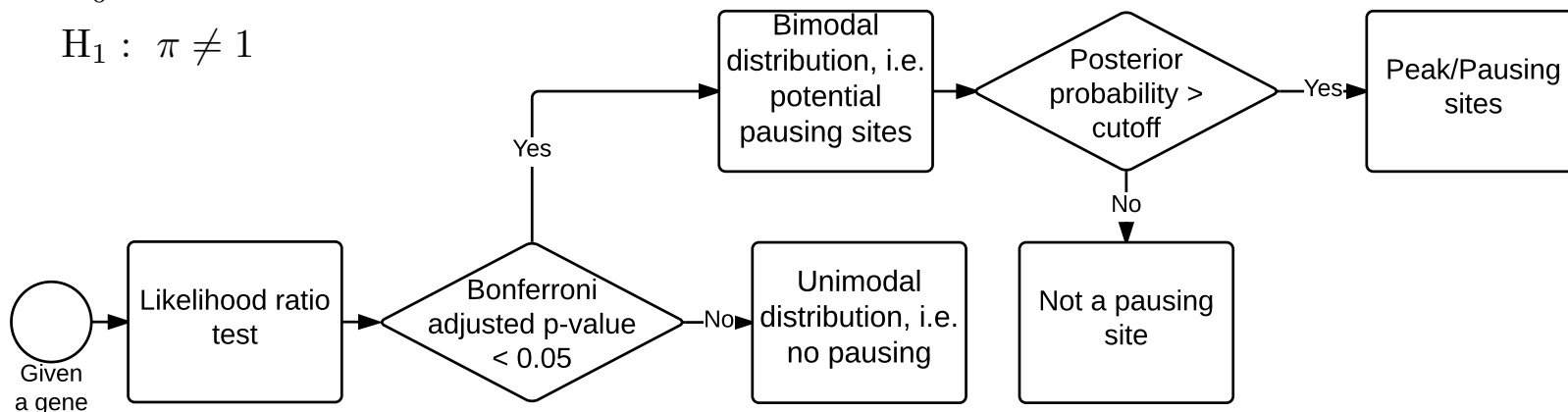
Yifei Huang

$$P(\mathbf{X}_i | \pi_i, \mu_i, k_i, r_i) = \prod_j \pi_i \mathcal{NB}(X_{ij} | \mu_i, r_i) + (1 - \pi_i) \mathcal{NB}(X_{ij} | k_i \mu_i, r_i),$$

for gene i at position j , where $k \geq 5$

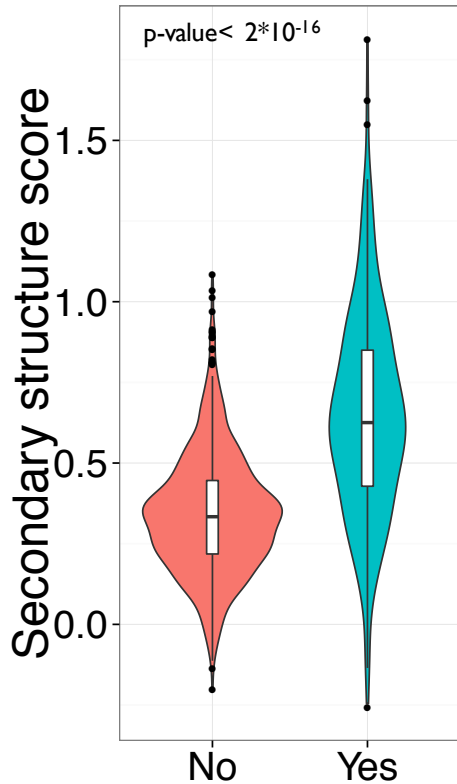
$H_0 : \pi = 1$

$H_1 : \pi \neq 1$

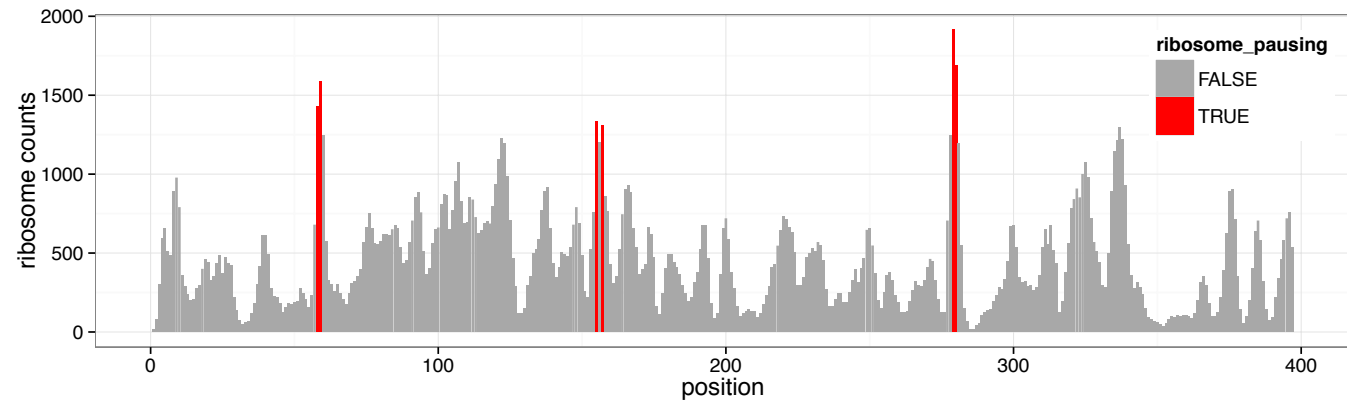
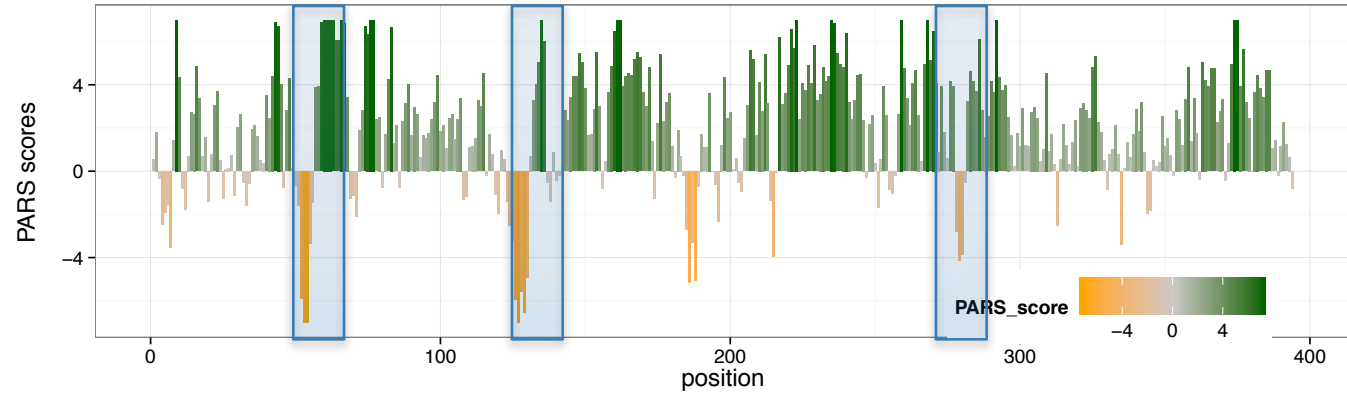
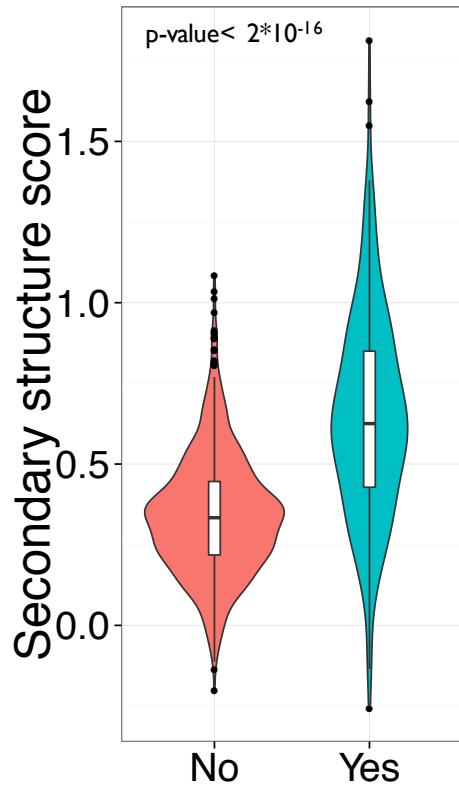


# genes	# genes (rpkm > 100)	# genes with pausing	# ribosome pausing sites identified
6664	1252	94	180

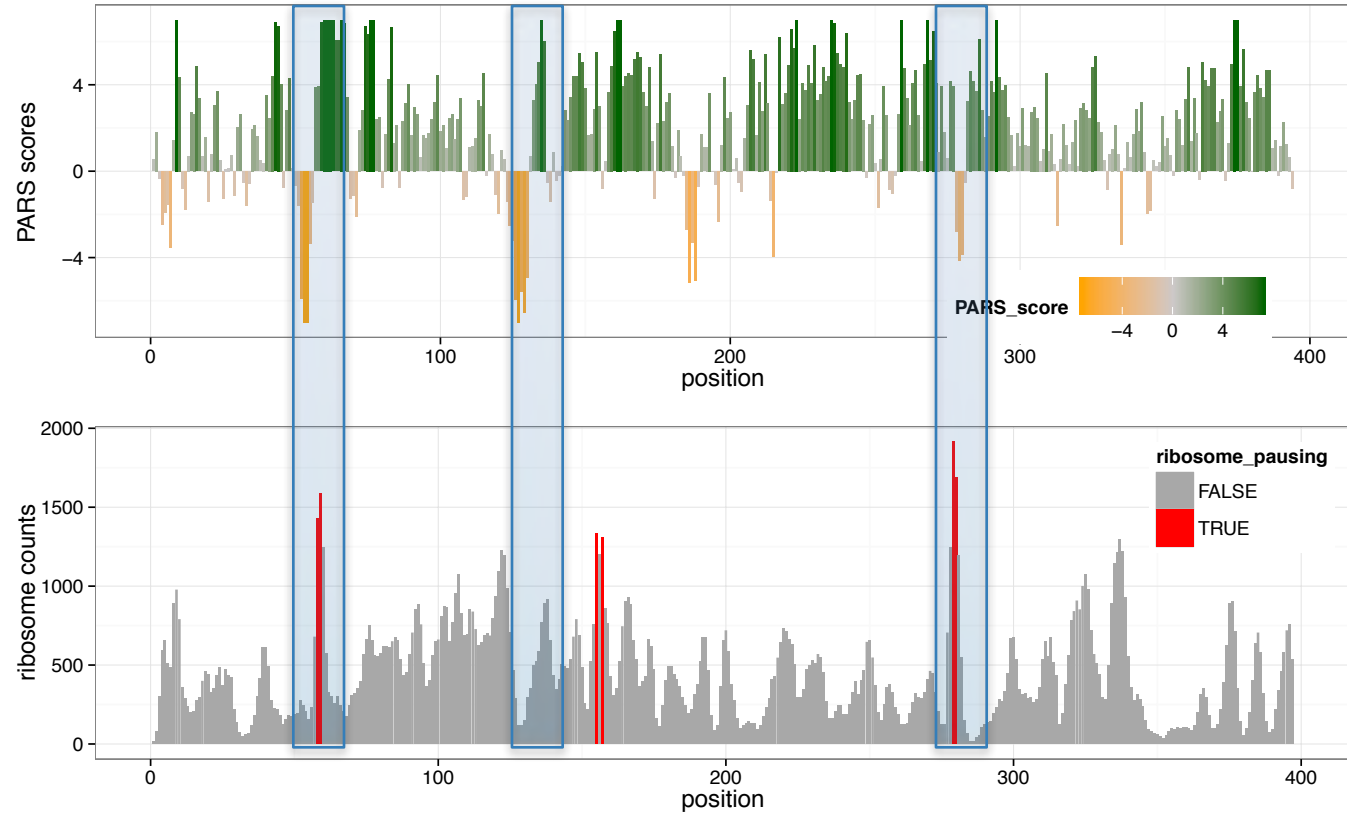
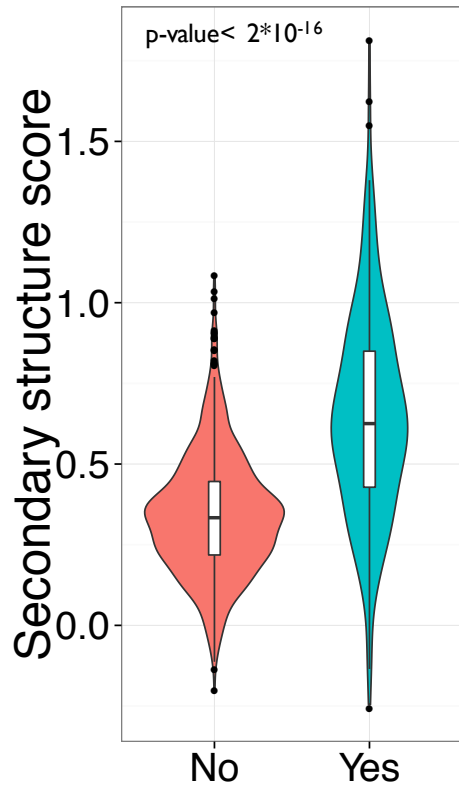
mRNA with stronger secondary structure tend to have ribosome pausing events



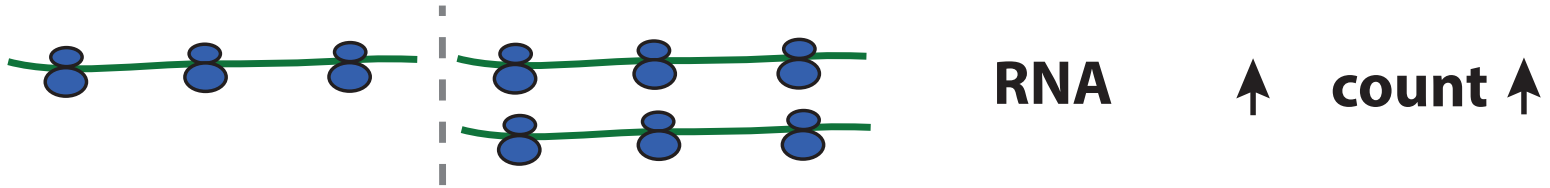
mRNA with stronger secondary structure tend to have ribosome pausing events



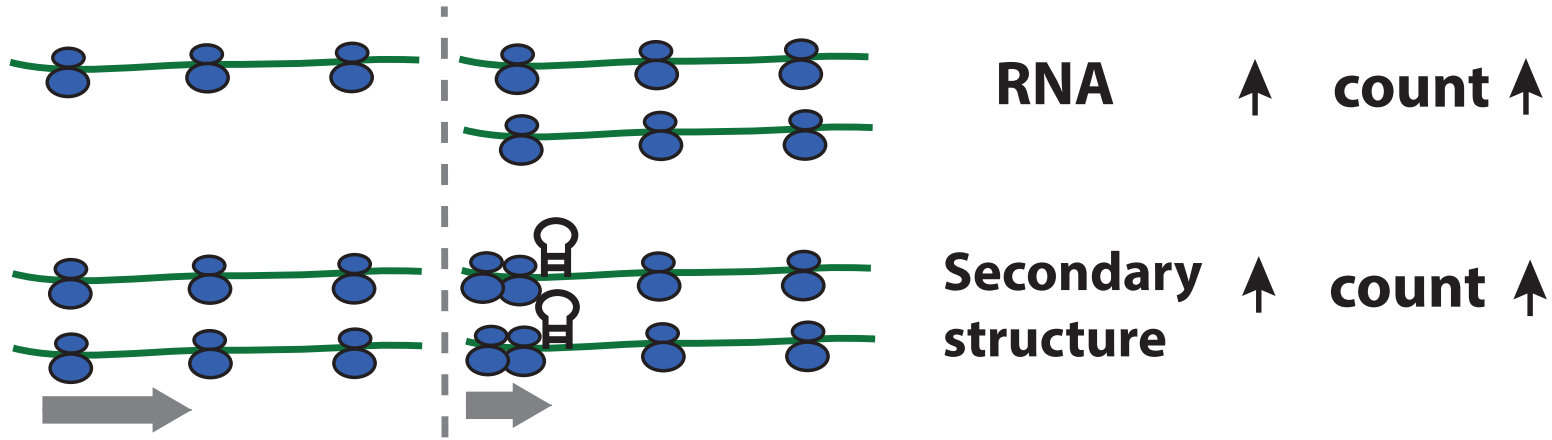
mRNA with stronger secondary structure tend to have ribosome pausing events



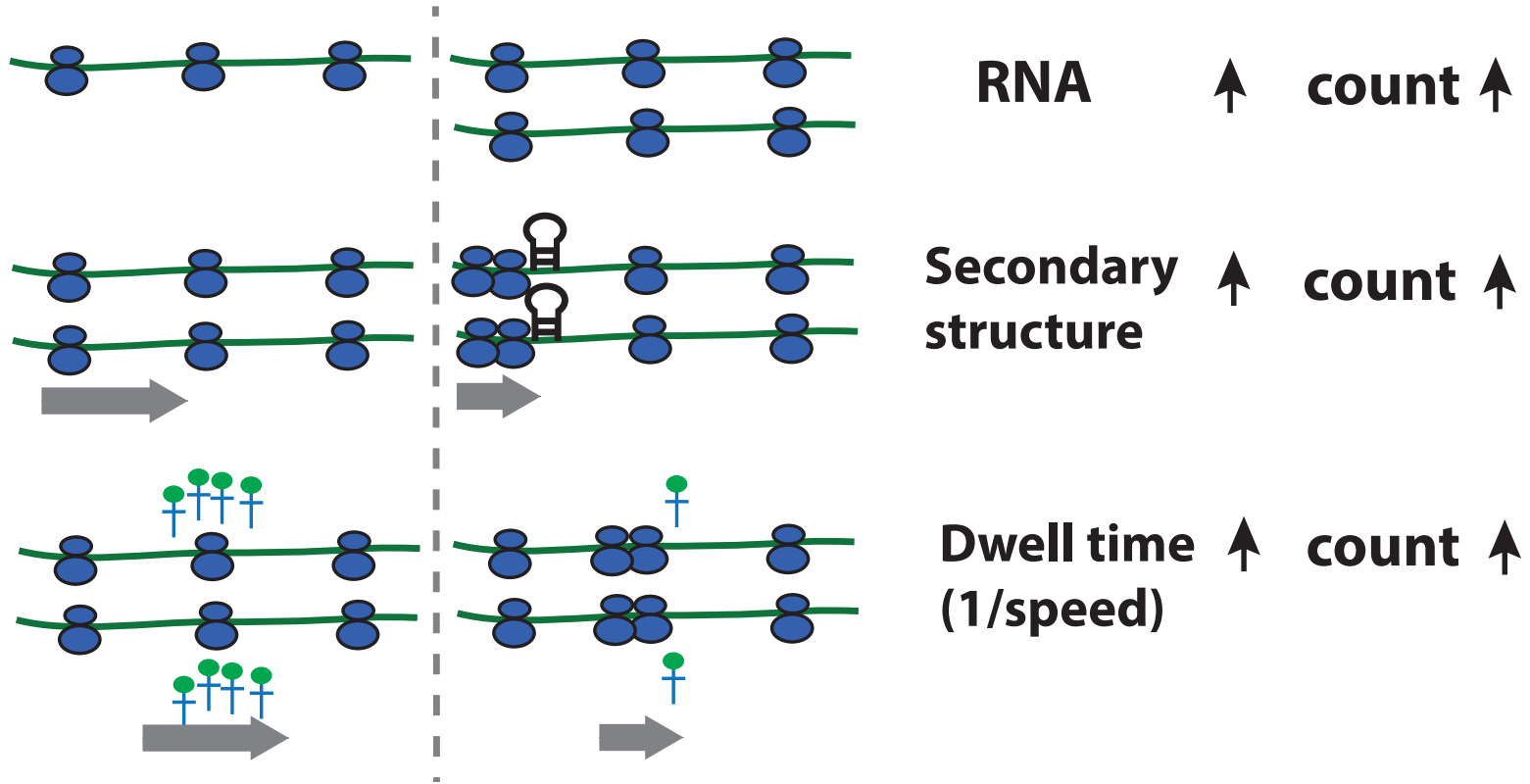
Simplifying the generalized linear model (GLM)



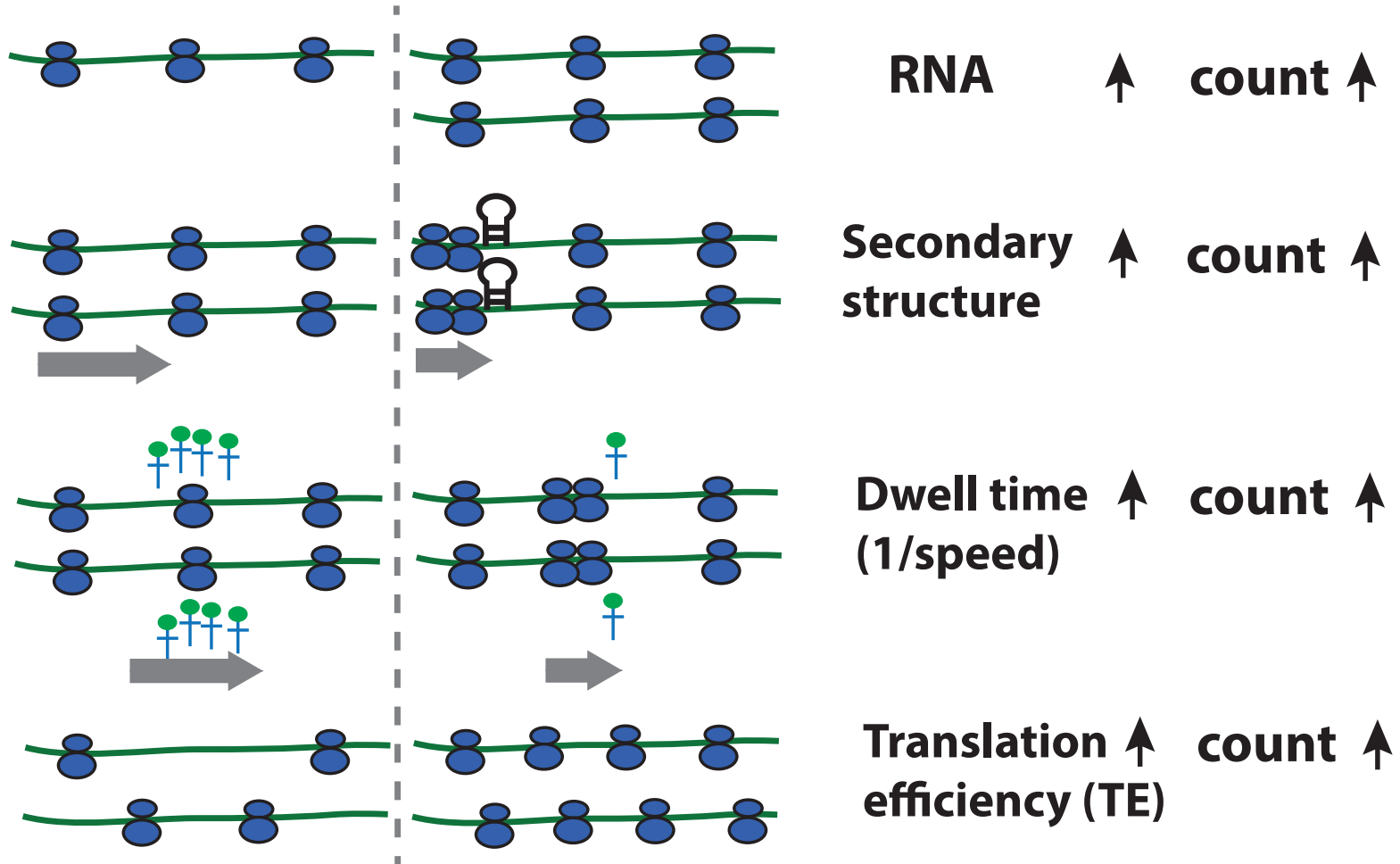
Simplifying the generalized linear model (GLM)



Simplifying the generalized linear model (GLM)



Simplifying the generalized linear model (GLM)



Joint inference of protein TE and codon dwell time using GLM, while accounting for secondary structure

GLM for joint inference of TE and codon dwell time:

$Y_{ij} \sim NB(\text{mean} = \mu_{ij}, \text{dispersion} = \alpha)$, for gene i , position j

$$g(\mu^{ij}) = \beta_0 + \underbrace{x_m^i}_{\text{mRNA}} + \underbrace{\beta_t^i}_{\text{TE}} + \underbrace{\beta_c^k}_{\text{codon}} + \underbrace{\beta_s x^{ij}}_{\text{secondary structure}}$$

where $g(\cdot)$ is a log link function, $\mu_{ij} = E(Y_{ij})$,

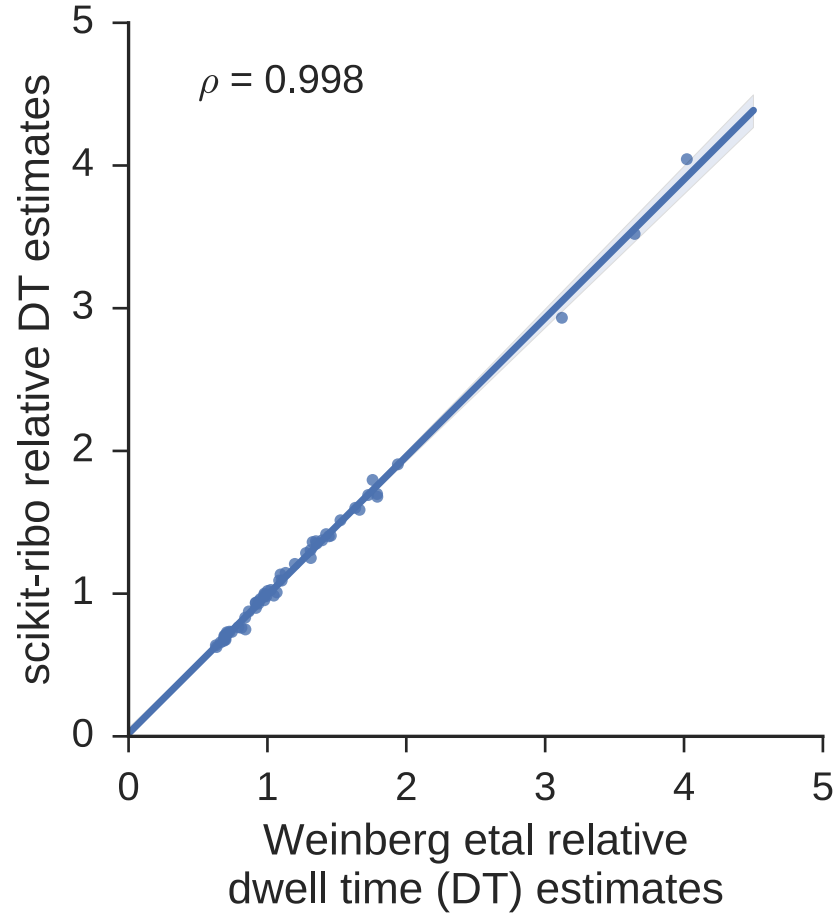
x_m^i is mRNA abundance for gene i ,

β_t^i is translational efficiency for gene i ,

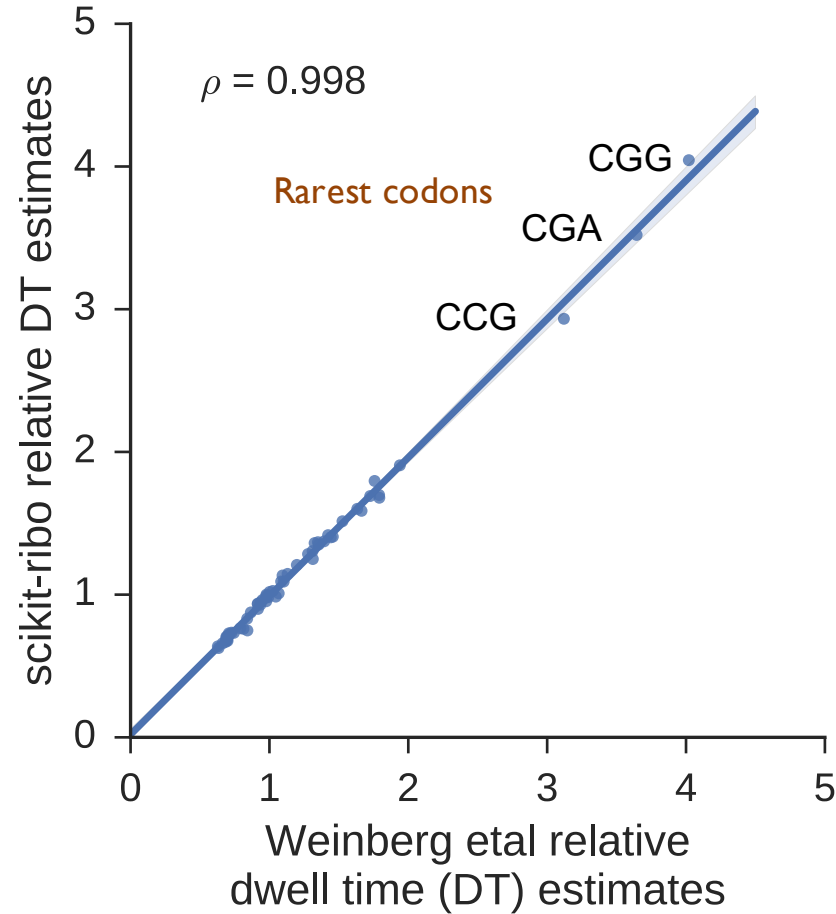
β_c^k is dwell time for codon k ,

$\beta_s x^{ij}$ is secondary structure effect at position j for gene i .

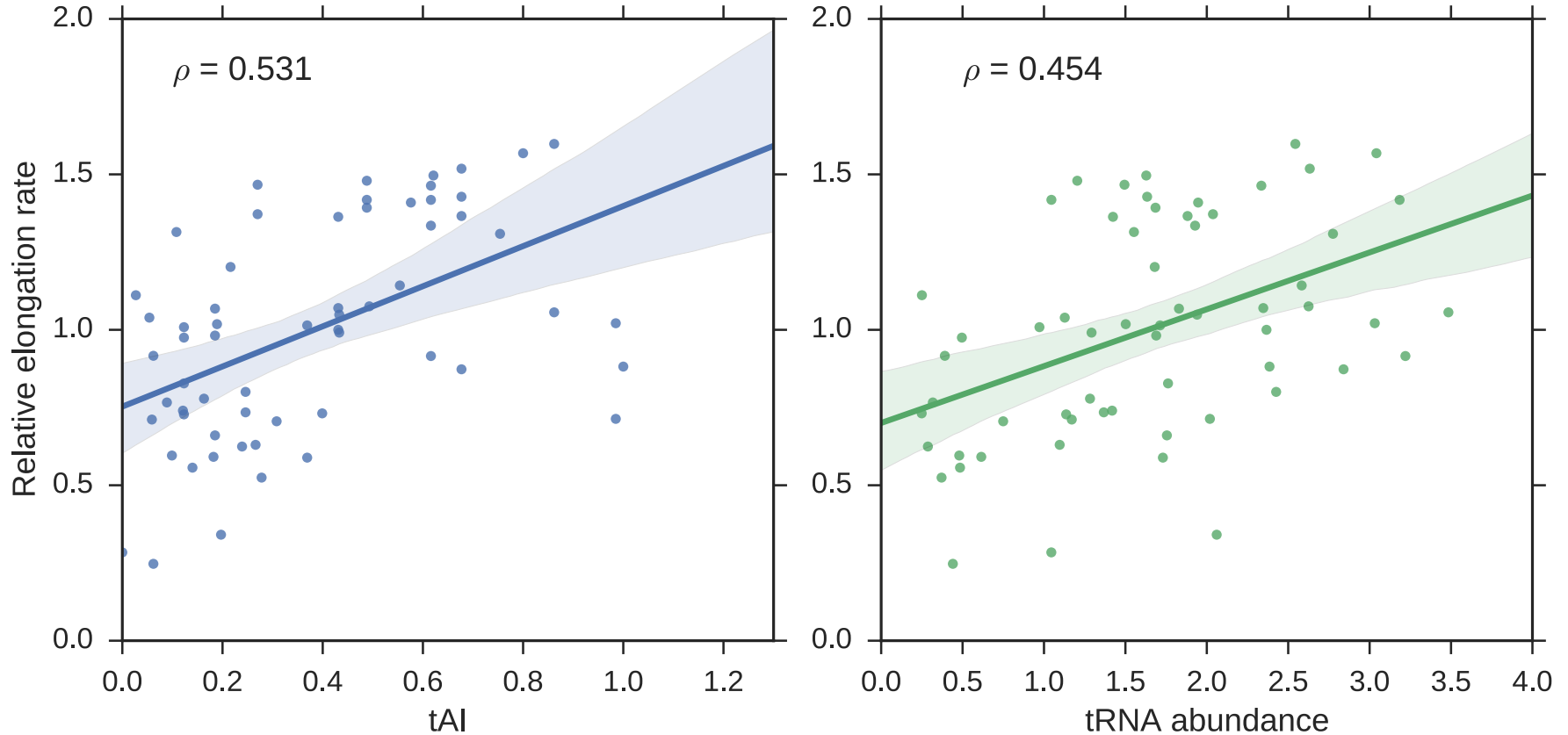
Scikit-ribo perfectly reproduced relative codon dwell time from Weinberg et al



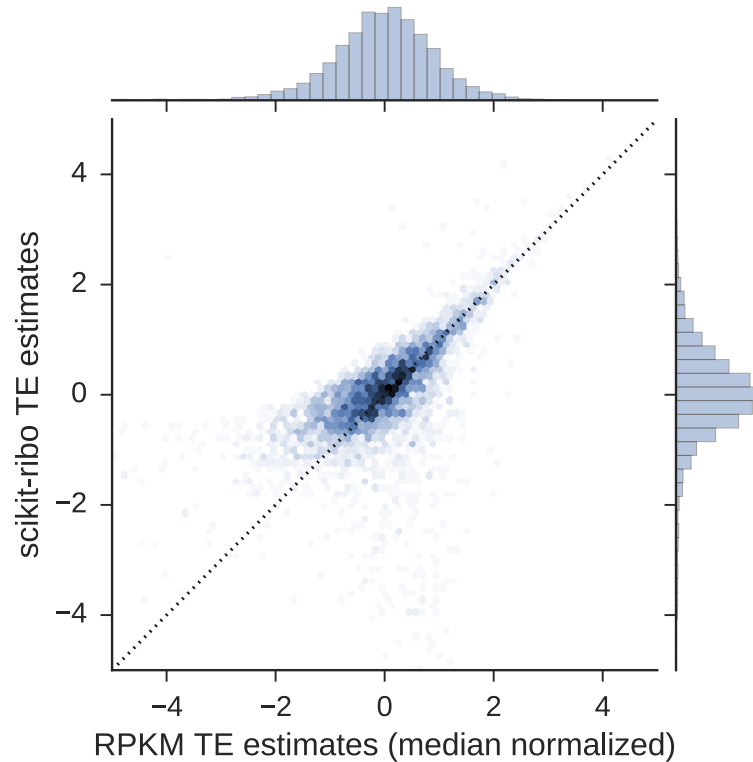
Scikit-ribo perfectly reproduced relative codon dwell time from Weinberg et al



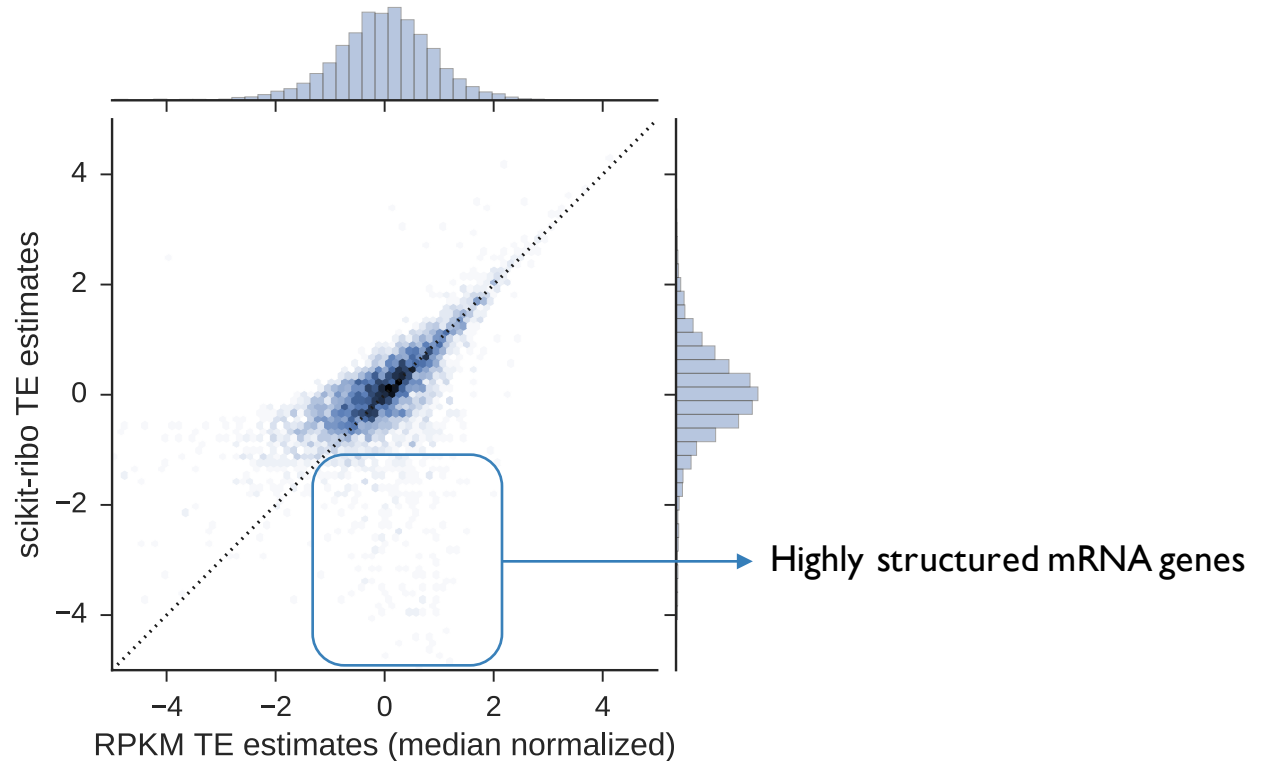
Significant correlation between tRNA abundance and codon elongation rates



GLM estimates vs. RPKM-based estimates reveals systematic bias in typical Riboseq analysis

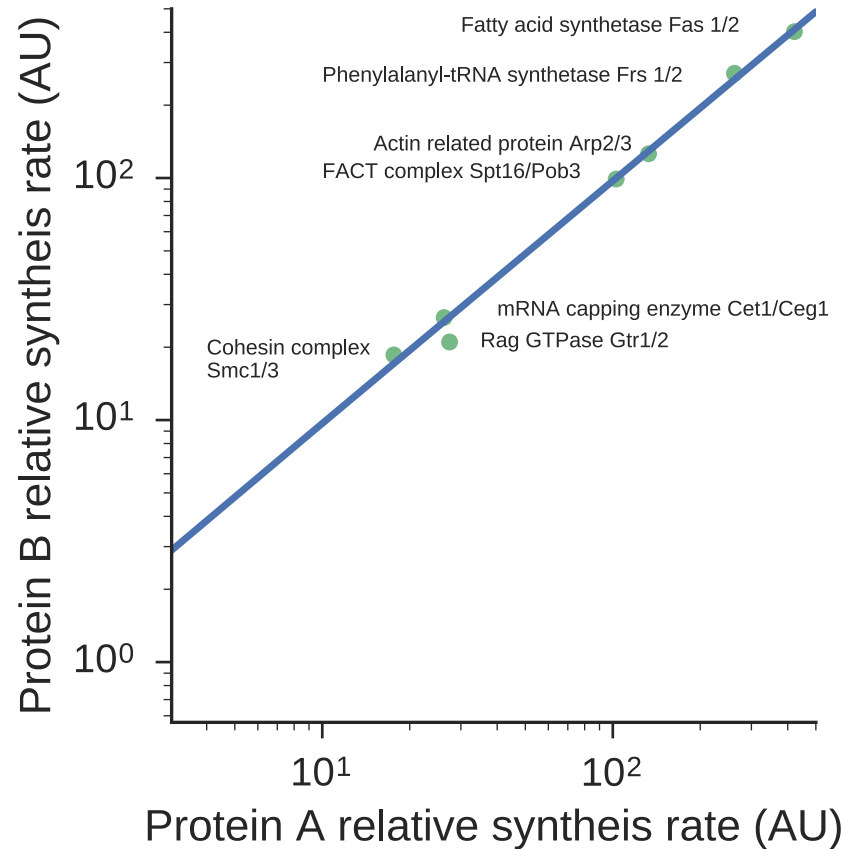


GLM estimates vs. RPKM-based estimates reveals systematic bias in typical Riboseq analysis



The rpkm based approach overestimated TE of highly structured mRNA, while the rest of the mRNA were slightly under-estimated, as hypothesized.

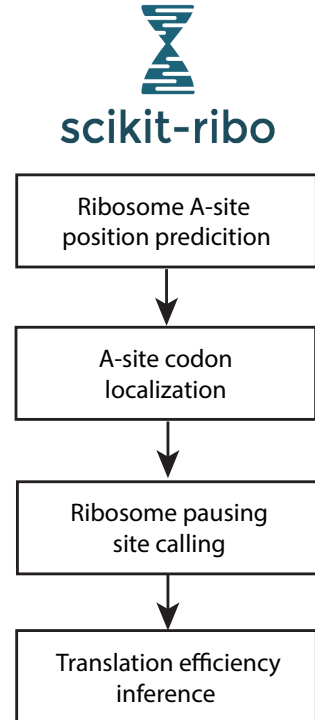
Accurate TE estimation supported by proportional synthesis for heterodimeric complexes in *S. cerevisiae*.



Summary

Discussed:

- 1) Introduced scikit-ribo for joint analysis of Riboseq and RNAseq data.
- 2) Identified biases in Riboseq data due to ribosome pausing.
- 3) Corrected biases and revealed underlying biology
- 4) Joint inference of codon elongation rate and protein TE
- 5) Revealed precise translational control at codon level



Acknowledgments

Lyon Lab

Max Doerfel
Yiyang Wu
Jonathan Crain
Jason O'Rawe



Gholson Lyon



Michael Schatz

Schatz Lab

Fritz Sedlazeck
Tyler Garvin
James Gurtowski
Maria Nattestad
Srividya Ramakrishnan

CSHL

Yifei Huang
Adam Siepel
Noah Dukler

JHU

Rachel Green

UCSF

Jonathan Weissman
Joshua Dunn
David Weinberg

Rutgers

Premal Shah

Stony Brook University

Rob Patro