Scikit-ribo - Accurate A-site prediction and robust modeling of translational control

Han Fang

October 29, 2015
Genome Informatics

Acknowledgments

Lyon Lab
Max Doerfel
Yiyang Wu
Jonathan Crain
Jason O'Rawe

Gholson Lyon

Michael Schatz

Schatz Lab
Fritz Sedlazeck
Tyler Garvin
Hayan Lee
James Gurtowski
Maria Nattestad
Srividya Ramakrishnan

Cold Spring Harbor Laboratory:

Yifei Huang
Noah Dukler
Melissa Kramer
Eric Antoniou
Elena Ghiban
Stephanie Muller

Stony Brook University:

Rob Patro

CSH Stanley Institute for Cognitive Genomics
Cold Spring Harbor Laboratory

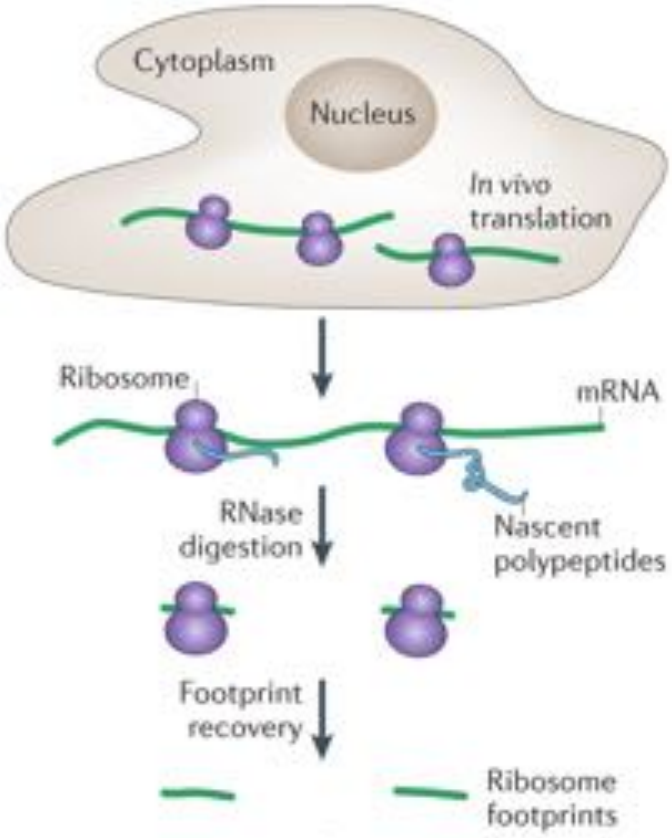
SFARI
Simons Foundation
Autism Research Initiative

NSF

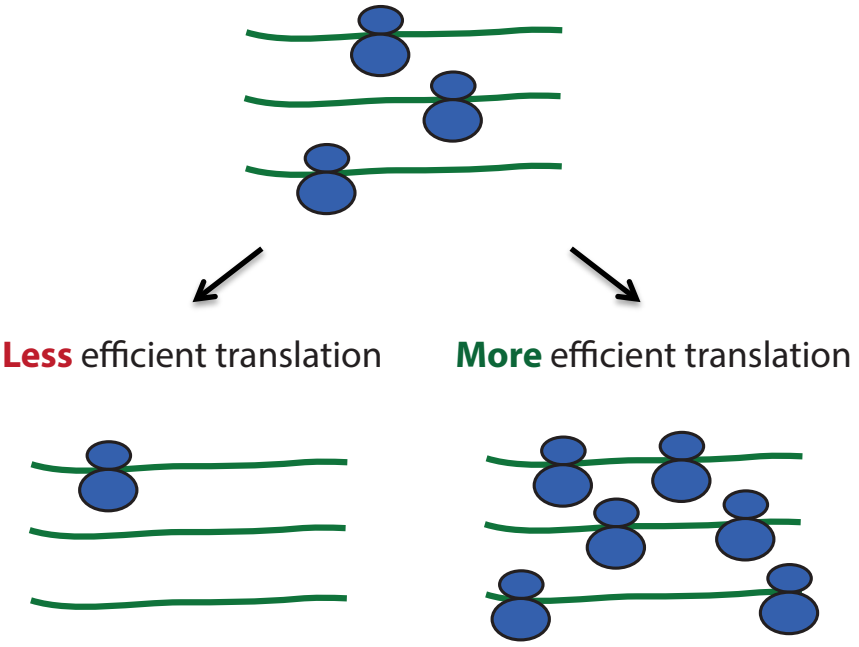
Stony Brook University

Central dogma of biology – Classic view

What is ribosome profiling (Riboseq)?



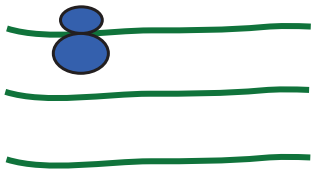
Normal translation efficieny (TE)



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

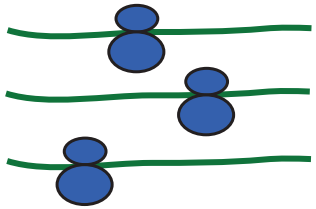
Calculate translational efficiency (TE)

Less efficient translation



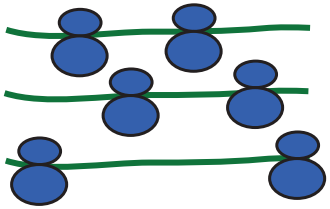
$\log_2(TE) < 0$

Normal translation efficieny (TE)



$\log_2(TE) = 0$

More efficient translation

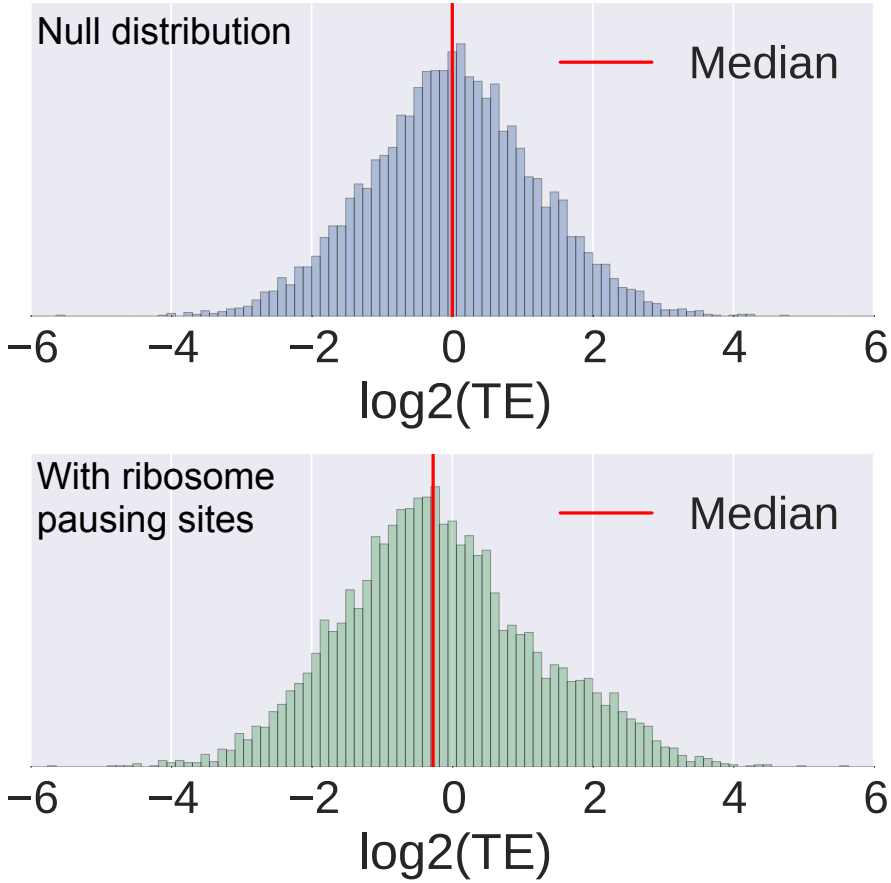


$\log_2(TE) > 0$

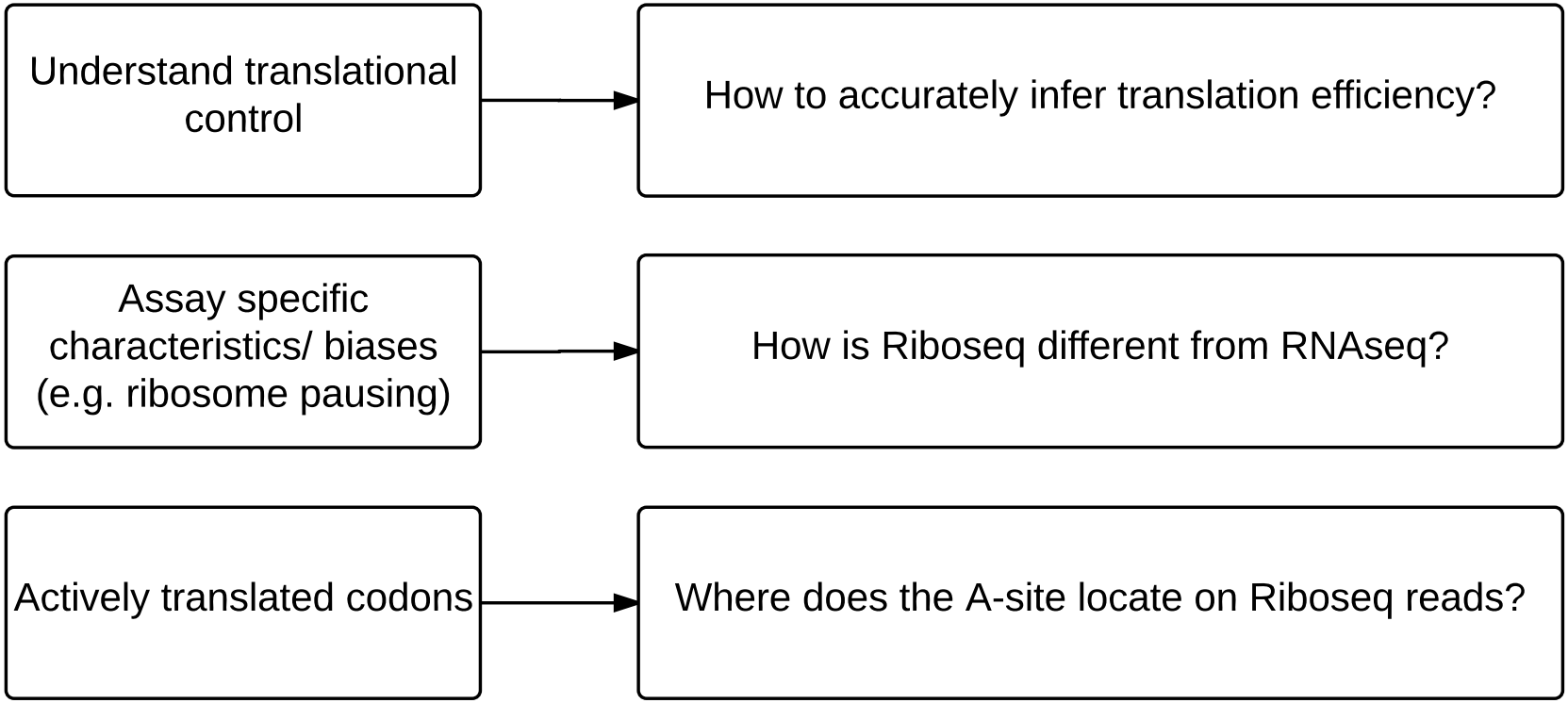
$$TE = \frac{Riboseq\ \ rpkm}{RNAseq\ \ rpkm}$$

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

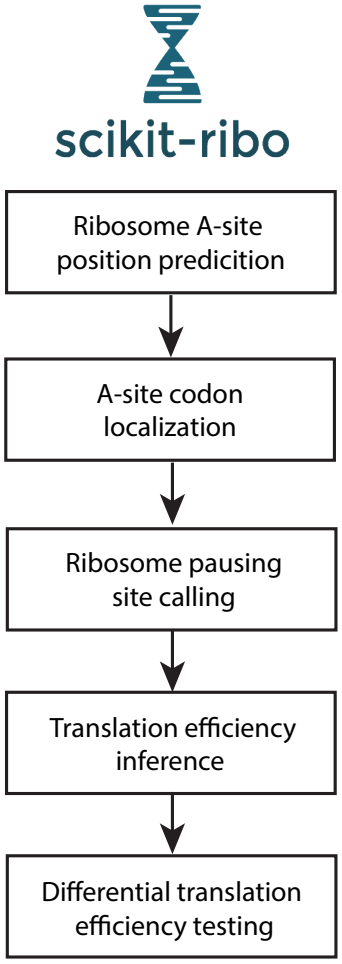
Simulated *S. cerevisiae* data - TE distribution are negatively-skewed by ribosome pausing events



Analytical Challenges



Introducing scikit-ribo



What and where is the ribosome A-site?

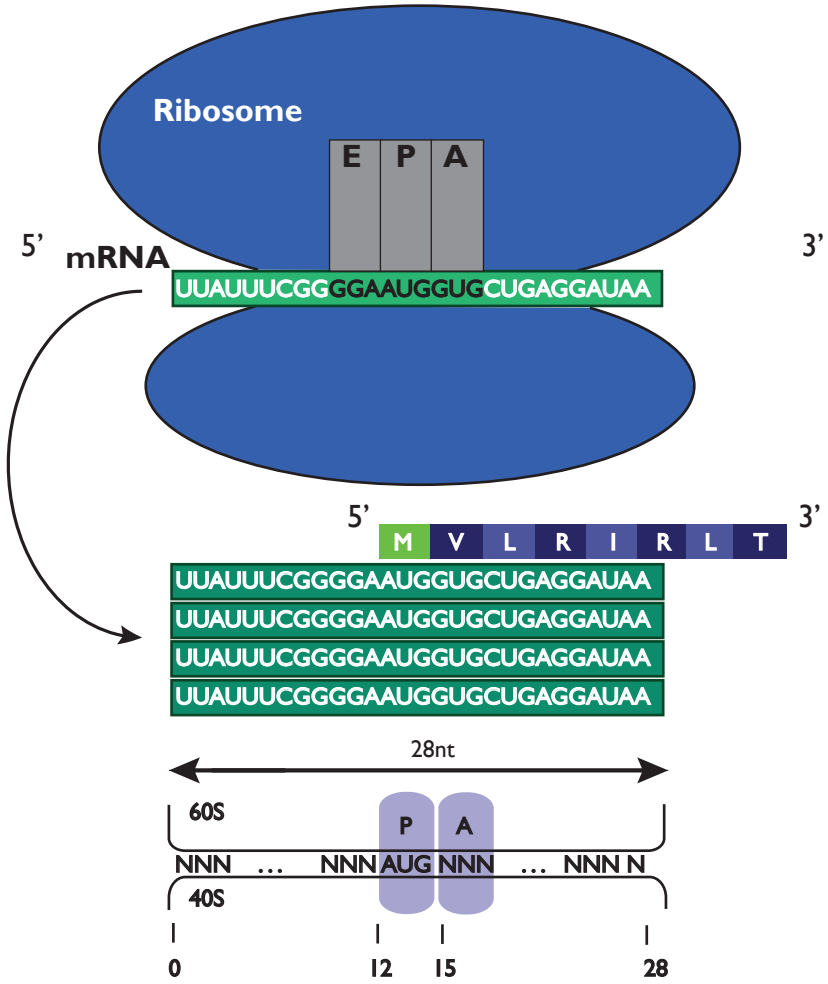
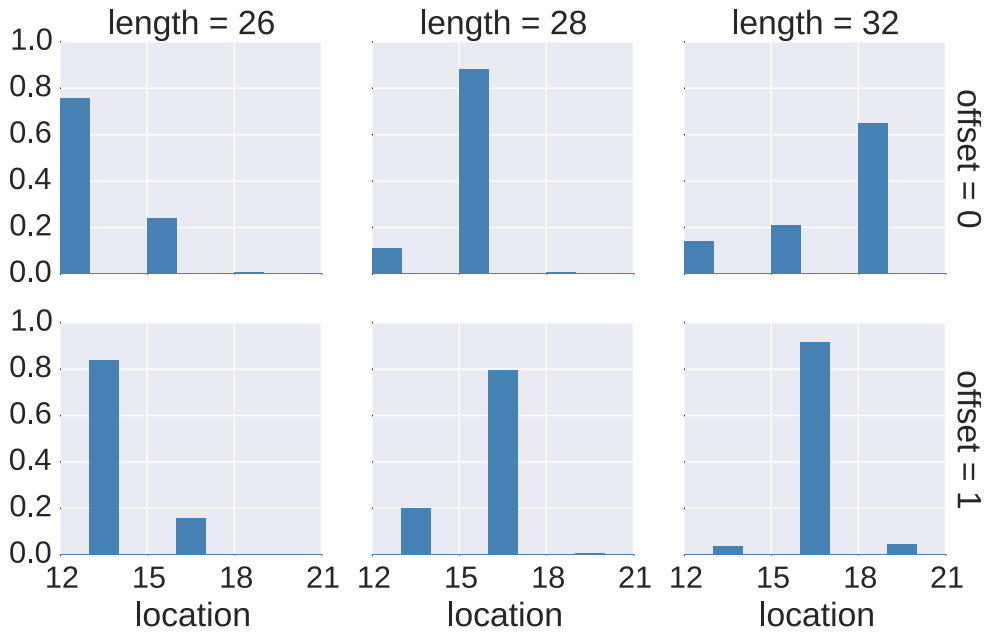
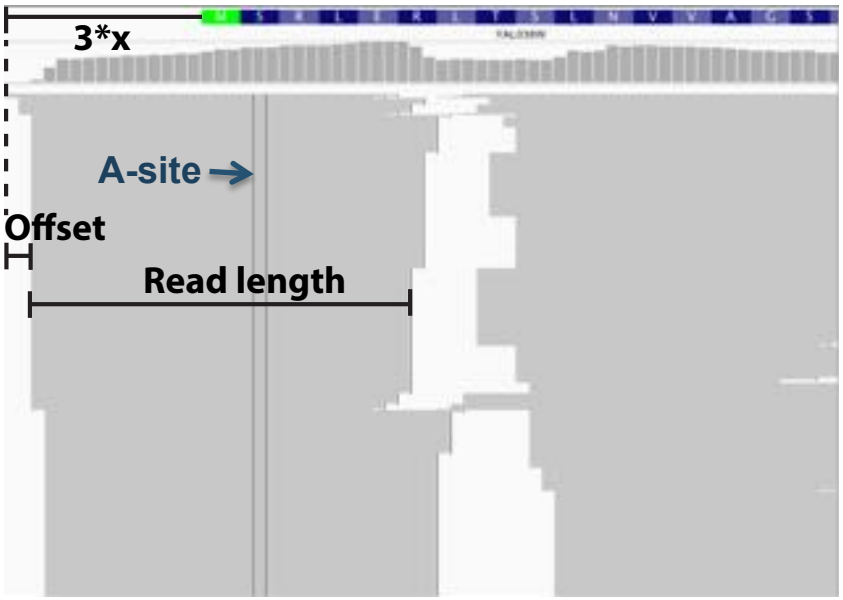


Figure adapted from Ingolia et al. *Science* (2009)



How to predict A-site?

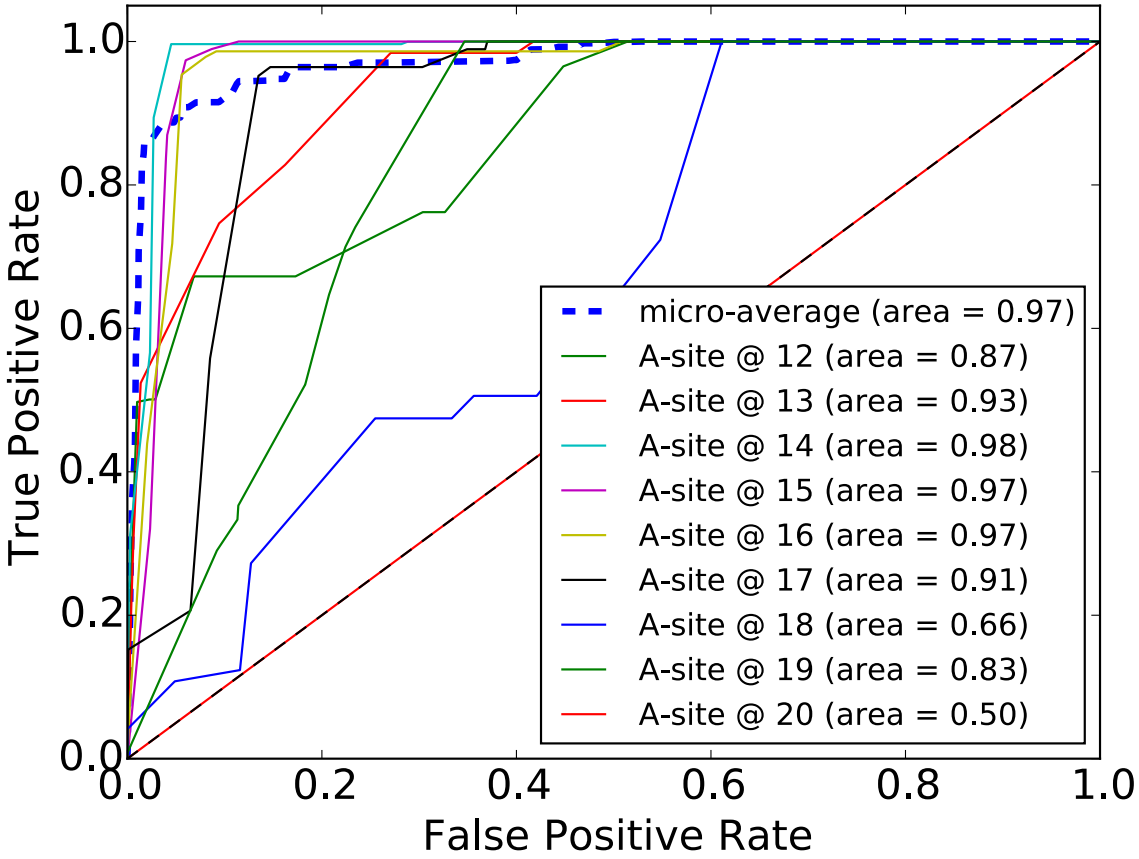
Training data and features:



Classifier and model tuning:

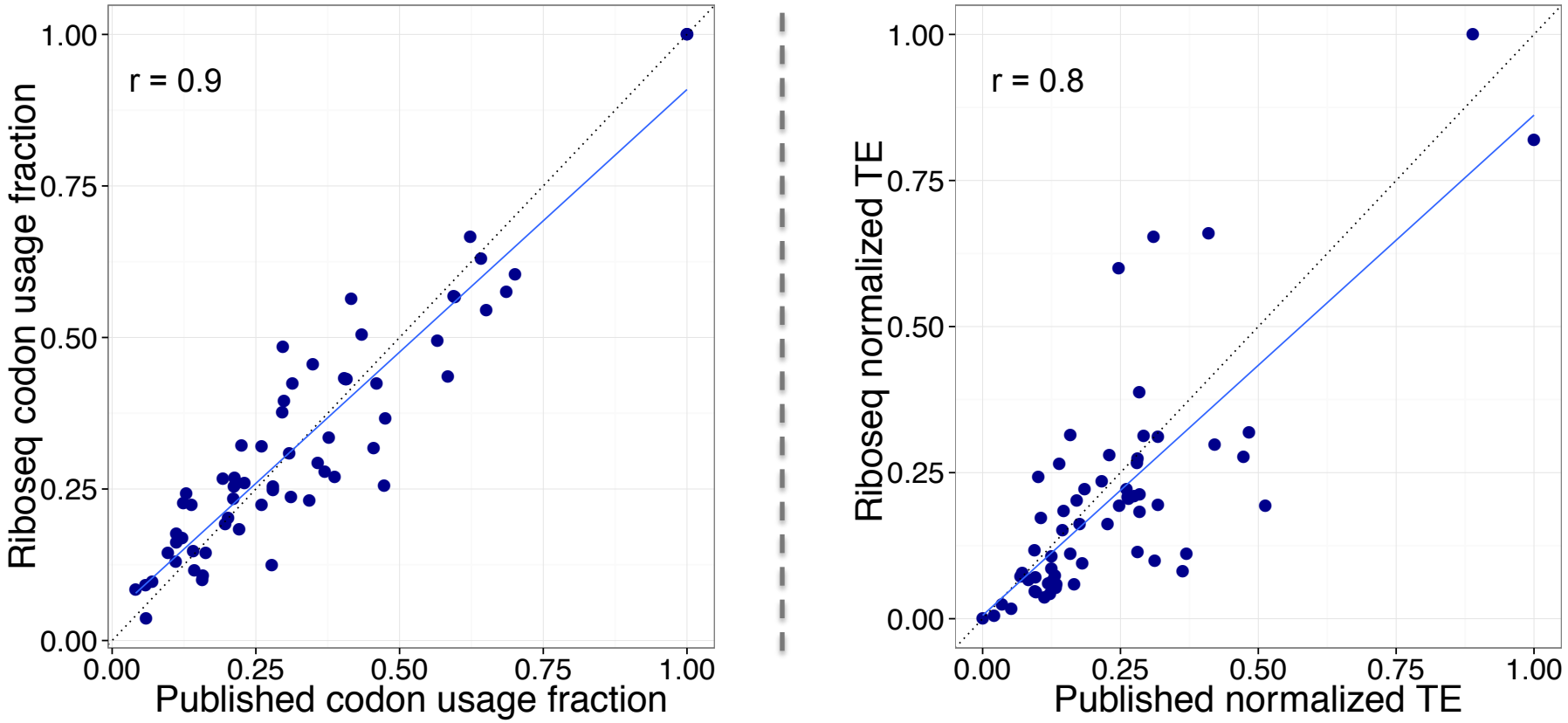
- SVM with RBF kernel (scikit-learn)
- 10 fold cross-validation for grid search
- Make predictions on all reads genome-wide

Prediction performance by cross validation



Scikit-ribo has much higher accuracy of identifying A-site than the previous method (0.86 vs. 0.64, 10-fold CV).

Scikit-ribo accurately predicted codon usage fraction and codon normalized TE

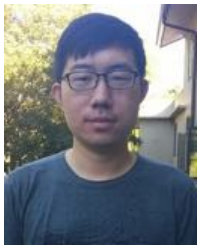


Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding, Pechmann, Frydman (2013).

Finding ribosome pausing sites (peaks) is hard. But it is easier after knowing the A-site location.

Q: how to robustly identify ribosome pausing sites while accounting for over-dispersion?

Ribosome pausing site identification by negative binomial mixture model



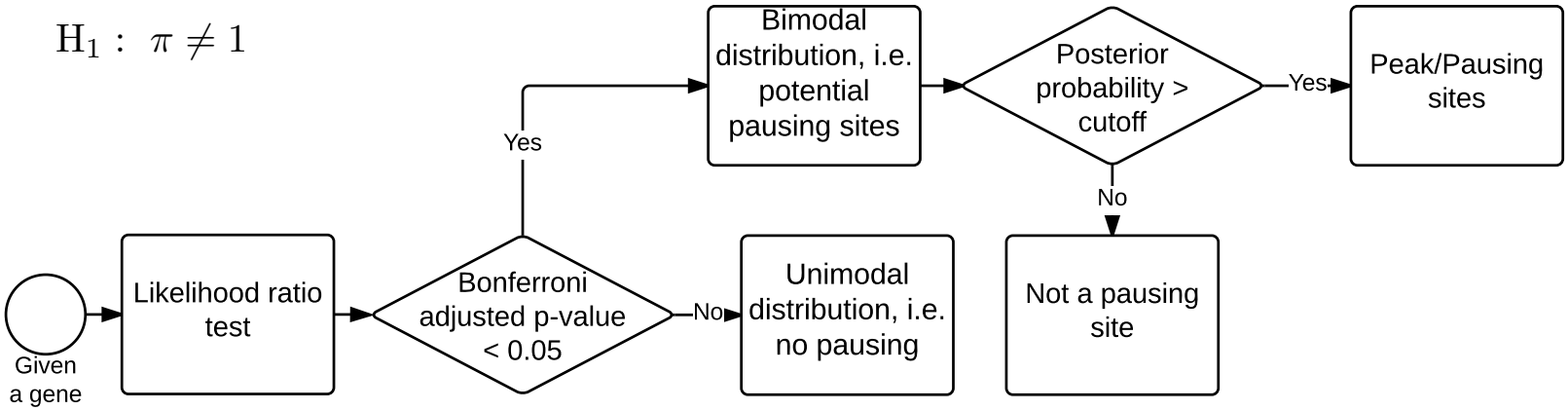
Yifei Huang

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

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

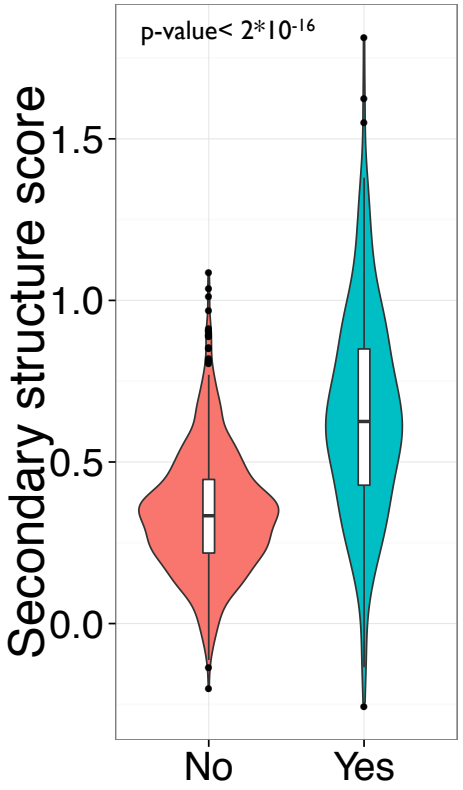
$H_0: \ \pi = 1$

$H_1: \ \pi \neq 1$



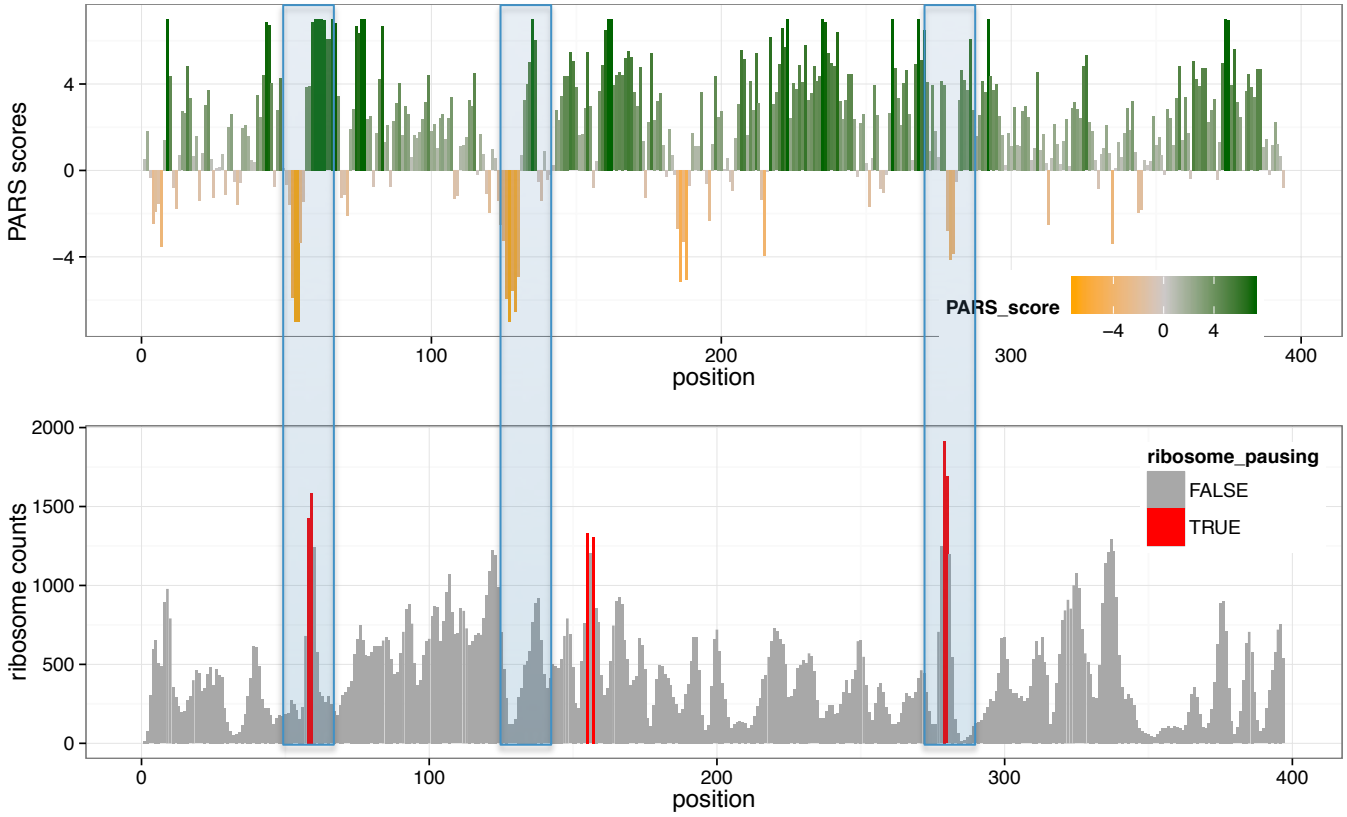
# genes	# genes (rpkm > 100)	# genes with pausing	# ribosome pausing sites identified
6664	1252	94	180

mRNA with stronger secondary structure tend to have ribosome pausing events

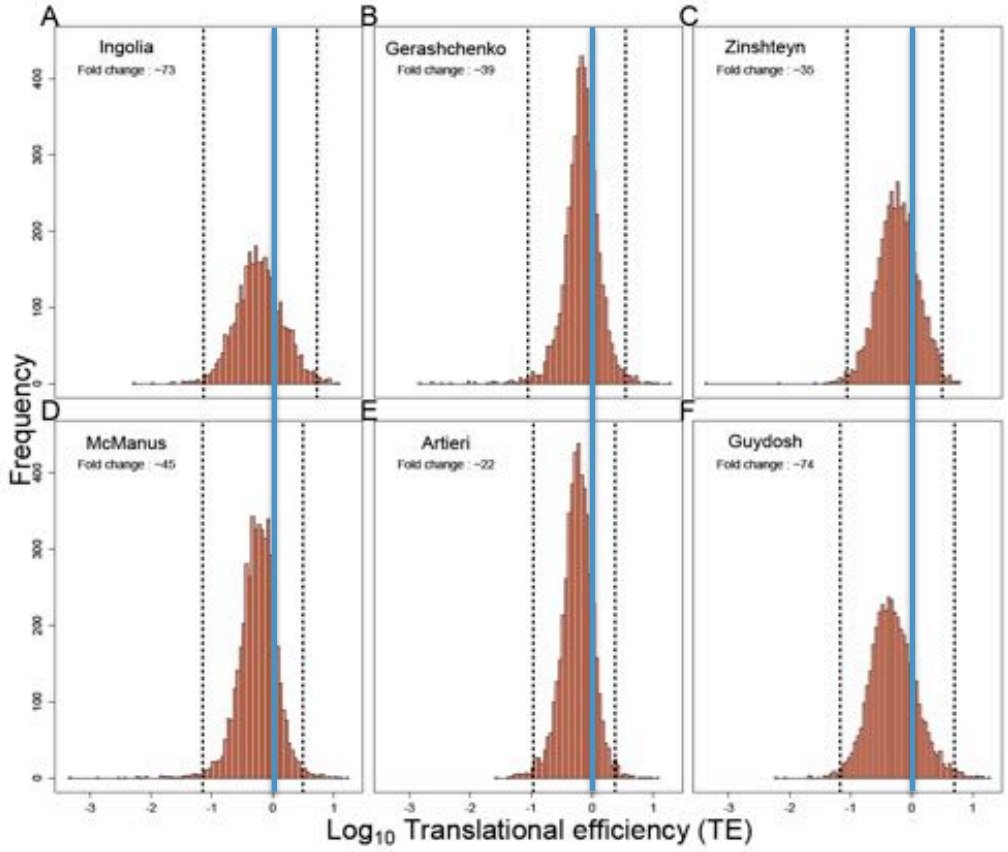


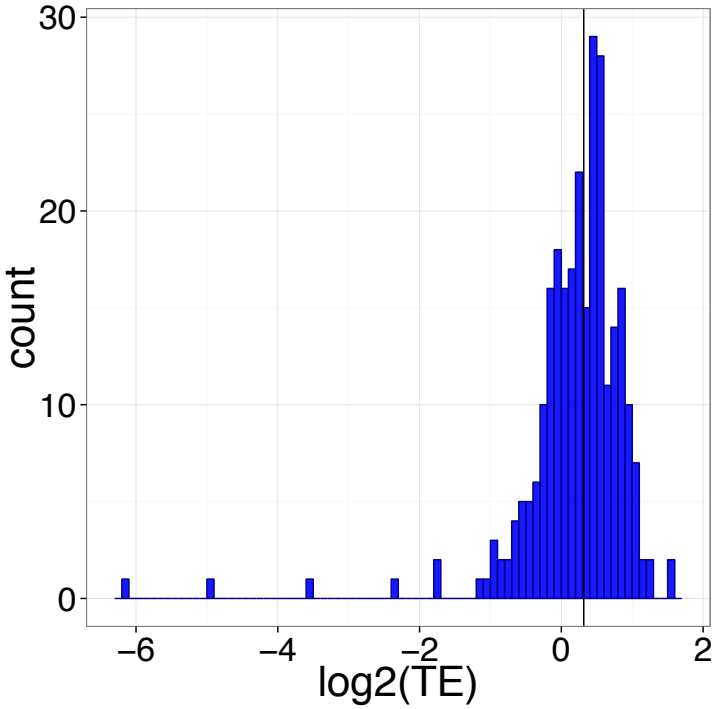
Fisher exact test p-value = 0.001

Kertesz et al. Nature (2010)



TE distributions are negatively-skewed in many studies. Over-structured mRNA show inflated TE.





Chi Square test p-value < $2*10^{-16}$

Improved ribosome-footprint and mRNA measurements provide insights into dynamics and regulation of yeast translation. Weinberg, Shah et al. (2015)

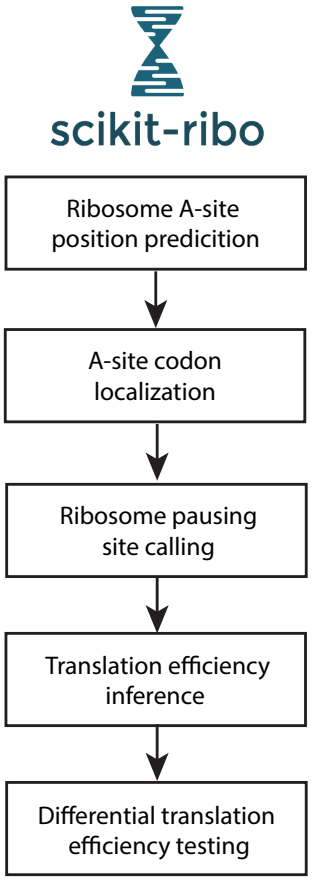
Summary

Discussed:

1) Introduce scikit-ribo for joint analysis of Riboseq & RNAseq data.
2) Learn from data itself to determine ribosome A-site location.
3) Reveal biases in Riboseq data due to ribosome pausing.
4) How Riboseq biases lead to issues with estimating TE.

Ongoing work:

1) Adjust for those biases and provide an unbiased estimate of TE.
2) Extend the ribosome pausing calling to a HMM based method.
3) Joint inference of translation initiation and elongation rates.



https://github.com/hanfang/