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

Han Fang
Acknowledgments

**Lyon Lab**
Max Doerfel
Yiyang Wu
Jonathan Crain
Jason O’Rawe

**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

---

[Images of researchers from Lyon Lab and Schatz Lab]
Central dogma of biology – Classic view

- **Replication**: DNA 
- **Transcription**: DNA to RNA (RNA Polymerase)
- **Translation**: RNA to Protein (Ribosome)
What is ribosome profiling (Riboseq)?

Calculate translational efficiency (TE)

<table>
<thead>
<tr>
<th>Less efficient translation</th>
<th>Normal translation efficiency (TE)</th>
<th>More efficient translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \log_2(TE) &lt; 0 )</td>
<td>( \log_2(TE) = 0 )</td>
<td>( \log_2(TE) &gt; 0 )</td>
</tr>
</tbody>
</table>

\[
TE = \frac{\text{Riboseq rpkms}}{\text{RNAseq rpkms}}
\]

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

- Understand translational control
- Assay specific characteristics/biases (e.g. ribosome pausing)
- Actively translated codons

Questions:
- How to accurately infer translation efficiency?
- How is Riboseq different from RNAseq?
- Where does the A-site locate on Riboseq reads?
Introducing scikit-ribo

- Ribosome A-site position prediction
- A-site codon localization
- Ribosome pausing site calling
- Translation efficiency inference
- Differential translation efficiency testing
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
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

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

\[ P(X_i | \pi_i, \mu_i, k_i, r_i) = \prod_j \pi_i \mathcal{NB}(X_{ij} | \mu_i, r_i) + (1 - \pi_i) \mathcal{NB}(X_{ij} | k_i \mu_i, r_i), \]

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

\( H_0 : \pi = 1 \)
\( H_1 : \pi \neq 1 \)

<table>
<thead>
<tr>
<th># genes</th>
<th># genes (rpkm &gt; 100)</th>
<th># genes with pausing</th>
<th># ribosome pausing sites identified</th>
</tr>
</thead>
<tbody>
<tr>
<td>6664</td>
<td>1252</td>
<td>94</td>
<td>180</td>
</tr>
</tbody>
</table>
mRNA with stronger secondary structure tend to have ribosome pausing events

Fisher exact test p-value = 0.001

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

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/