Supplementary Material

SeqFold: Genome-scale reconstruction of RNA secondary structure integrating high-throughput sequencing data

Zhengqing Ouyang, Michael P. Snyder, and Howard Y. Chang

Supplementary Figure 1 The correspondence between the accessibility and structure preference. The median accessibilities and 95% bootstrap confidence intervals over the SS, UN, and DS sites are shown. The median accessibilities and 95% confidence intervals from 1,000 datasets with shuffled structure preferences are shown as a control.

Supplementary Figure 2 Application of SeqFold to FragSeq data (Underwood et al. 2010). (a) The known secondary structure of U1a, the RNAstructure MFE prediction (middle, sensitivity = 0.44 and PPV = 0.42), and the SeqFold prediction (right, sensitivity = 0.52 and PPV = 0.58). (b) The known secondary structure of U3b, the RNAstructure MFE prediction (middle, sensitivity = 0.95 and PPV = 0.85), and the SeqFold prediction (right, sensitivity = 0.97 and PPV = 0.80). (c) The known secondary structure of U5, the RNAstructure MFE prediction (middle, sensitivity = 1 and PPV = 0.86), and the SeqFold prediction (right, sensitivity = 1 and PPV = 0.91). Note that the improvement with FragSeq data is small presumably because it is sparser and only has information on single stranded regions.

Supplementary Figure 3 (a) The predicted ribosome density matches the measured ribosome density. Ribosome density is predicted from a linear regression model with RNA accessibility, codon bias and mRNA expression as predictors. The percent of explained variance ($R^2$) in ribosome density is indicated. a.u., arbitrary unit. (b) The regression coefficients of global RNA accessibility and codon bias for predicting ribosome density. Average RNA accessibility is calculated in the [-40, +40] base window around translation start site. Codon bias (CAI) is measured from the entire coding region. (c) The regression coefficients of local RNA accessibility and codon bias for predicting
ribosome density. RNA accessibility and codon bias (CAI) are calculated based on a non-overlapping sliding window of 3 bases.

**Supplementary Figure 4** The P values of anticorrelations between various histone marks and 5’ end RNA accessibilities calculated from RNAfold in a 20-bp-wide sliding window. The data points are the centers of the windows.

**Supplementary Figure 5** The distributions of motif counts in the true targets as well as the false targets of the RBP Puf3. Many RBP unbound transcripts contain the same motif consensus as the bound transcripts do. Only those transcripts with at least one motif instance are included.

**Supplementary Figure 6** The average accessibilities along motif positions for the true targets and false targets of RBPs. In most cases the average accessibilities of the motifs of true targets are higher than those of the false targets. Above each subplot are the name of the RBP and the motif consensus. Puf3 and Vts1 each has two motif consensuses which are both included.

**Supplementary Figure 7** Incorporating accessibility for distinguishing true and false targets of RBPs other than those in Fig. 5. ROC curves are shown for predictions using motif count only and predictions incorporating RNA accessibility. Above each subplot are the name of the RBP and the motif consensus.

**Supplementary Figure 8** Improvement of target identification using SeqFold derived accessibility versus RNAfold derived accessibility on four RBPs. ROC curves are shown for predictions using motif count only, incorporating SeqFold-derived RNA accessibility, and incorporating RNAfold-derived RNA accessibility. Above each subplot are the name of the RBP and the motif consensus.

**Supplementary Table 1** SeqFold reconstructed RNA secondary structures for over 3,000 yeast transcripts with PARS data.
**Supplementary Table 2** Comparison of motif count and SeqFold-derived accessibility for distinguishing true and false RBP targets.
Relative contribution of RNA secondary structure on translation efficiency

To assess the relative contribution of RNA secondary structure on determining translation efficiency, we applied predictive modeling for predicting ribosome densities in combination with mRNA expression and codon bias. There has been much debate on whether codon bias or RNA secondary structure is the main sequence determinant of translation efficiency (Kudla et al. 2009; Tuller et al. 2010). We used a linear model to evaluate their relative contribution by including both features in the same model framework. The average RNA accessibility in the [-40, +40] base region around the translation start site was calculated for each yeast transcript. Codon adaption index (Sharp and Li 1987) (CAI), a measurement of codon bias, was calculated from the entire coding region of each transcript. Expression level of mRNAs, measured from RNA-Seq experiment (Nagalakshmi et al. 2008), was also put in the model as it is correlated with ribosome density (Ingolia et al. 2009) or protein abundance (Lu et al. 2007; Vogel et al. 2010). The transcript-level ribosome density \( y \) (log-scale), mRNA expression \( e \) (log-scale), codon bias \( c \) (log-scale), and average RNA accessibility \( s \) were scaled to have mean 0 and standard deviation 1. The linear model for predicting ribosome density is as the formula below

\[
y = \beta_e e + \beta_c c + \beta_s s + \varepsilon
\]

The model was fitted by ordinary least squares on ~2,000 yeast transcripts and the explained variance \( R^2 \) was calculated. Together, all these factors explained 71.2% of ribosome density variation between the RNAs (Supplementary Fig. 3a). To assess the relative importance of RNA secondary structure versus codon bias, we examined the regression coefficients in the model, which quantifies the number of units that ribosome density is expected to increase as a predictor increases by one unit, controlling for all other predictors. As shown in Supplementary Fig. 3b, although the global codon bias of the entire coding region is more strongly associated with ribosome binding than the average accessibility of the RNA, both contributions are statistically significant. In fact, adding RNA accessibility as the third variable is statistically significant \( (P = 3 \times 10^{-11}, \text{likelihood ratio test}) \), improving the \( R^2 \) from 0.706 to 0.712.
We further investigated how local (in contrast to global) compositions of codon bias and RNA secondary structure affect ribosome density of a transcript. We fitted a series of linear models based on the features in a non-overlapping sliding window of size 3, the length of a codon, along the transcripts by the formula below:

\[ y = \beta_i^c e + \beta_i^l c^l + \beta_i^s s^l + \epsilon^l \]

where \( i \in \ldots, -6, -3, 3, 6, \ldots \) indicating the distance of the windows to the start codon. \( c^l \) and \( s^l \) were calculated from the triplets that were covered by the windows. Strikingly, the regression coefficients of local RNA accessibility in the upstream region of translation start site were almost comparable to those of local codon bias in the coding region (Supplementary Fig. 3c). In conclusion, the relative contribution of RNA secondary structure on translation efficiency was reinforced with genome-scale study of local variation patterns.
Supplementary Figure 1

Accessibility

Measured
Shuffled

Structure preference

DS UN SS
Supplementary Figure 2

a

U1a

b

U3b

c

U5

Known

RNAstructure MFE

SeqFold with FragSeq
Supplementary Figure 3

(a) Measured vs. predicted ribosome density (a. u.)

(b) Regression coefficient for ribosome density prediction

(c) Position relative to start codon

$R^2 = 0.712$

$P = 3E-11$

$P = 2E-153$
Supplementary Figure 4

![Graph showing the value of anticorrelation with RNAfold-derived accessibility across different positions downstream of TSS. The x-axis represents position downstream of TSS, ranging from 0 to 100, and the y-axis represents the p-value of anticorrelation, ranging from 1e-05 to 1e-01. Different lines represent various histone modifications and proteins such as H3, H4, H3K9ac, H3K14ac, H4ac, H3K4me3, ESA1, GCN5, and their positions downstream of TSS vary in the graph.](image-url)
Supplementary Figure 5

- Motif count vs. Number of transcripts
- Targets: True, False
- Y-axis: Number of transcripts
- X-axis: Motif count
- Bars indicate distribution of transcripts with different motif counts.
Supplementary Figure 6

The figure consists of multiple subplots, each comparing the effectiveness of RNA binding for true and false targets at various positions. The x-axis represents the position, while the y-axis shows the averaged accessibility. Each subplot includes a line graph with two lines: one for true targets and another for false targets.

**Subplots:**
- **Khd1 HWNCUUWY**
- **Msl5 UACUAAC**
- **Nab2 DRARAMGMD**
- **Pub1 HUUUJUHW**
- **Pu2 UAAAUW**
- **Pu3 CNJUAAUA**
- **Pu4 UGAJUWJJA**
- **Pu5 WUGJWUUW**
- **Vts1 CNGG**
- **Vts1 CNGGN**
- **Y1932J AAYACCCY**
- **Pu3 UGAUAAUA**

**Legend:**
- Red line: True targets
- Blue line: False targets
Supplementary Figure 7

- Khd1 HWNCAUWY
- Nab2 DRARAMGMD
- Pub1 HUUUUUHW
- Puf2 UAUAAUW
- Vts1 CNGG
- Vts1 CNGGN
Supplementary Figure 8

Ouyang et al., Supplementary p. 13
**Supplementary Table 1** See the supplementary Excel file.

**Supplementary Table 2** Comparison of motif count and SeqFold-derived accessibility for distinguishing true and false RBP targets.

<table>
<thead>
<tr>
<th>RBP</th>
<th>Source</th>
<th>Motif consensus</th>
<th>Count AUC</th>
<th>Accessibility AUC</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Khd1</td>
<td>Hogan</td>
<td>HWNCAUUWY</td>
<td>0.614</td>
<td>0.659</td>
<td>7.3%</td>
</tr>
<tr>
<td>Msl5</td>
<td>Berglund</td>
<td>UACUAAC</td>
<td>0.566</td>
<td>0.684</td>
<td>20.8%</td>
</tr>
<tr>
<td>Nab2</td>
<td>Hogan</td>
<td>DRARAMGMD</td>
<td>0.591</td>
<td>0.639</td>
<td>8.1%</td>
</tr>
<tr>
<td>Pub1</td>
<td>Hogan</td>
<td>HUUUUUUHW</td>
<td>0.636</td>
<td>0.672</td>
<td>5.7%</td>
</tr>
<tr>
<td>Puf2</td>
<td>Hogan</td>
<td>UAAUAUW</td>
<td>0.645</td>
<td>0.637</td>
<td>-1.2%</td>
</tr>
<tr>
<td>Puf3</td>
<td>Gerber</td>
<td>CNUGUAHAUA</td>
<td>0.520</td>
<td>0.758</td>
<td>45.8%</td>
</tr>
<tr>
<td>Puf3</td>
<td>Gerber</td>
<td>UGUHAUA</td>
<td>0.587</td>
<td>0.732</td>
<td>24.7%</td>
</tr>
<tr>
<td>Puf4</td>
<td>Hogan</td>
<td>UGUAHMNUA</td>
<td>0.539</td>
<td>0.658</td>
<td>22.1%</td>
</tr>
<tr>
<td>Puf5</td>
<td>Hogan</td>
<td>WUUGUAWUWU</td>
<td>0.550</td>
<td>0.629</td>
<td>14.4%</td>
</tr>
<tr>
<td>Vts1</td>
<td>Aviv</td>
<td>CNGG</td>
<td>0.699</td>
<td>0.701</td>
<td>0.3%</td>
</tr>
<tr>
<td>Vts1</td>
<td>Aviv</td>
<td>CNGGN</td>
<td>0.699</td>
<td>0.704</td>
<td>0.7%</td>
</tr>
<tr>
<td>Yll032c</td>
<td>Hogan</td>
<td>AAUACCY</td>
<td>0.553</td>
<td>0.658</td>
<td>19.0%</td>
</tr>
</tbody>
</table>

AUC, area under curve of ROC plot.
References