Quantitative model suggests both intrinsic and contextual features contribute to the transcript coding ability determination in cells

Yu-Jian Kang[1], Jing-Yi Li[1], Lan Ke[1], Shuai Jiang[1], De-Chang Yang[1], Mei Hou[1], Ge Gao[1,*]

[1] Biomedical Pioneering Innovation Center (BIOPIC), Beijing Advanced Innovation Center for Genomics (ICG), Center for Bioinformatics (CBI), and State Key Laboratory of Protein and Plant Gene Research at School of Life Sciences, Peking University, Beijing, 100871, China

* To whom correspondence should be addressed. Tel: +86-010-62755206; Email: gaog@mail.cbi.pku.edu.cn

Biographical Note:

Yu-Jian Kang is a Postdoctoral researcher at Peking University. She is interested in developing novel computational methods to model gene expression regulation *in silico*.

Jing-Yi Li is a Ph.D. candidate at Peking University. He is interested in developing deep learning methods for omics data.

Lan Ke is a PhD candidate at Peking University. She is interested in using computational methods to understand functionality of long noncoding RNAs (lncRNAs).

Shuai Jiang is a PhD candidate at Peking University. She is interested in delineating the lncRNA landscape effectively and efficiently.

De-Chang Yang is a PhD candidate at Peking University. He is interested in developing novel computational methods to annotate the function and evolution of lncRNAs.

Mei Hou is a PhD candidate at Peking University. She is interested in designing and implementing online tools for understanding transcriptome.

Ge Gao is a Principal Investigator at the Biomedical Pioneering Innovation Center (BIOPIC) & Beijing Advanced Innovation Center for Genomics (ICG), Peking University. His research group focuses on developing novel computational tools for deciphering the function and evolution of gene regulation circuits.

Abstract

Gene transcription and protein translation are two key steps of the "*central dogma*". It is still a major challenge to quantitatively deconvolute factors contributing to the coding ability of transcripts in mammals. Here, we propose Ribosome Calculator (RiboCalc) for quantitatively modeling the coding ability of RNAs in human genome. In addition to effectively predicting the experimentally confirmed coding abundance via sequence and transcription features with high accuracy, RiboCalc provides interpretable parameters with biological information. Large-scale analysis further revealed a number of transcripts with a variety of coding ability for distinct types of cells (i.e., context-dependent coding transcripts, CDCTs), suggesting that, contrary to conventional wisdom, a transcript's coding ability should be modeled as a continuous spectrum with a context-dependent nature.

Introduction

Gene transcription and protein translation are two key steps of the "*central dogma*". While protein abundance is generally believed to be regulated by both transcriptional and translational control [1-3], it is still a major challenge to quantitatively factors contributing to transcript's coding ability (i.e., whether a particular transcript will encode a protein and, if so, the corresponding abundance).

Benefitting from rapid development on high-throughput technology recently, several quantitative models have been proposed for modeling coding ability *in silico* based on various features in unicellular organisms [4-7]. While these models are rather accurate (e.g. the correlation of with ribosomal density has achieved 0.68[7]), heterogeneity across cells and species hinders their application in depicting translation control in mammals [8].

Multiple translation-related signatures have been reported in human and other mammal systems, revealing several gene-encoded transcription and translation regulatory features which substantially contribute to the final mRNA and protein expression levels [9-15]. Along this line, Volkova *et al.* have assessed these features and build qualitative models to discriminate coding and noncoding RNAs, as well as high- and low-translated mRNAs [16]. Trösemeier *et al.* have introduced a codon-specific translation elongation model to simulate ribosome dynamics during mRNA translation and integrate model's parameters for protein expression prediction [17].

Here we present an experiment-backed, data-oriented computational model (named Ribosome Calculator, RiboCalc) for quantitatively predicting the coding ability (Ribo-seq expression level) of a particular human transcript (Figure 1A). Features

collected for RiboCalc model are biologically related to translation control. We build the model using linear regression with Lasso penalty so that the feature parameters are easily connected to their contribution to transcript coding process. Multiple evaluations show that RiboCalc not only makes quantitatively accurate predictions but also offers insight for sequence and transcription features contributing to transcript coding ability determination, shedding lights on bridging the gap between the transcriptome and proteome. All scripts and data are available online at https://github.com/gao-lab/RiboCalc/.

Materials and Methods

Ribosome profiling data collection

We retrieved human data published since 2012 from RPFdb[18] and the NCBI BioProject database[19] by searching for the terms “ribosome profiling”, “ribosome profile”, “Ribo-seq”, “ribosome footprint” and “RPF”. We then manually selected Ribo-seq samples with paired RNA-seq data and without treatment interfering with translation. As a result, 61 datasets were retained from 30 studies covering 22 different human tissues or cell lines (Supplementary Table 1). The pipelines of transcriptome analysis for RNA-seq and Ribo-seq data are at https://github.com/gao-lab/RiboCalc/blob/master/feature_calculation/RNAandRibo-seq_processing.txt.

Mass spectrometry (MS) data analysis

To build a reliable coding ability prediction model, the first step is to identify *bona fide* coding transcripts, especially given several recent reports that a lot of annotated noncoding RNAs (ncRNAs) were found to encode peptides [20-22]. We selected a Ribo-seq based coding gene identification method which covered more than 90% of MS-based callings (Supplementary Figure 1, Supplementary Table 2 and 3).

To find the criteria for coding gene identification covering most MS observations, we compared the ribosome profiling based results with the MS results. As reported, mass spectrometry identifies proteins with high specificity but limited sensitivity[23]. Therefore, taken MS results as golden positive calls, we selected criteria to ensure that Ribo-based results covered 90% MS calls while with less false positives. The MS dataset for assessment was PXD002395 from the PRIDE database [24]. The overlapping cell lines in this MS project and our ribosome profiling data were HEK 293, HeLa and U-2 OS cells, so only MS data from these 3 cell lines were analyzed (Supplementary Table 2).

We then compiled a human protein database by adding theoretically translated peptides of transcript putative ORFs and protein-coding transcript translation sequences annotated by GENCODE release 24. Redundant protein sequences with identity higher than 90% were trimmed using CD-hit[25] in the database. We used pFind3 as the search engine [26]. 286 common contaminant proteins were automatically added into the original protein database by pFind. The reverse protein sequences were used as a decoy database for false discovery rate (FDR) control. The search parameters were as follows: 1) Trypsin/P digestion. 2) The precursor tolerance and fragment tolerance were set to 20 ppm and 20 ppm, respectively. 3) The search included variable modifications of methionine oxidation and N-terminal acetylation. 4) Minimal peptide length was set to six amino acids and a maximum of two missed cleavages was allowed. Peptides were filtered with a FDR threshold of 1%. We identified a gene as coding in MS when supported by at least one protein with a pFind Q-value less than 0.01 and more than 5 peptide fragments.

The approaches that we used for translated ORF scanning in ribosome profiling data were RiboCode[27] and ribORF[28]. Genes with translated ORF were identified as coding in Ribo-seq. We compared the MS and ribosome profiling results in corresponding cell lines (Supplementary Table 3). MS-based coding genes were taken as positive calls. Abbreviations in the equations below are as follows: FN, false negative; FP, false positive; TN, true negative; and TP, true positive.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}; \ \text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

RiboCode showed a lack of consistency with MS analysis (Supplementary Table 3), while ribORF, with a p-value cutoff of approximately 0.5, could achieve a sensitivity of 90% (Supplementary Figure 1). Thus, we adopted ribORF with a p-value higher

than 0.5 as an approach for translated ORF identification.

Transcript's coding ability contributes to the protein abundance determination process significantly. During the last two decades, the most common method for large-scale experimental determination of transcripts' coding ability is to measure the abundance of the corresponding proteins based on MS methods. Nevertheless, MS-based protein identification is less sensitive to short peptides and peptides expressed at low levels[29]. In addition, MS analyses often adopt a database-dependent search strategy that ignores genomic mutations and RNA editing events, hampering identification of the complete protein pool[30]. Meanwhile, by sequencing the RNA fragments protected by ribosomes, Ribo-seq measures translational activity in a quantitative manner with base resolution. Hence, Ribo-seq derived transcript density is an appropriate measure for evaluation and estimation of coding ability. During ribosome profiling data processing, we used MS results as the gold standard for determining the threshold of Ribo-seq based methods. This strategy ensured reliable, sensitive and precise identification of MS-supported protein-coding genes in the Ribo-seq data, providing an accurate training set for model building.

Identification of translated ORFs

After scanning translated ORF based on ribosome profiling data, we made pairwise comparisons between overlapping ORFs to select a most likely translated one. Translated ORFs were identified according to the following procedure (Supplementary Figure 2): Scenario 1) overlapping translated ORFs with unique regions covered by ribo-reads were both retained. Scenario 2) the ORFs without unique ribo-reads covering region were filtered out when their overlapping ORFs had. Scenario 3) if the unique regions of both overlapping ORFs were not covered by ribo-reads, the shorter ORF was left. Thus, we compiled a non-redundant catalog of translated ORFs. Since the translated ORFs defined by us were based on ribosome

profiling data, they might differ from the main ORFs in canonical annotation (Supplementary Figure 3). If there were no ribosome evidences (such as independent testing without Ribo-seq data), the longest ORF was taken as the putative translated ORF.

Calling transcript's coding status

As we scanned translated ORFs in Ribo-seq, we identified coding transcripts (have translated ORF) for each sample. Then, a transcript would be called as "noncoding" in a particular sample only if 1) not covered by any ribo-reads, 2) its expression abundance is higher than the lower bound of called "coding" ones in this sample. The expression threshold was set to the 300th quantile of transcripts per million (TPM) for all coding transcripts in the corresponding sample, so, at this abundance, translated ORFs could be effectively detected. By using the 300th quantile as a threshold, a large number of expressed transcripts (96,968) were removed (Supplementary Figure 4), indicating that the process is stringent enough for ncRNA identification. Another circumstance is that a transcript is covered by Ribo-seq reads but no translated ORF is identified. We also removed this kind of transcripts (1,031) because of lacking reliable evidence for coding.

As described above, we identified coding and noncoding transcripts in each sample. Transcripts with translated ORF in every expressed sample were classified as coding, while the transcripts consistently without translation were classified as noncoding. In addition, several transcripts showed coding in some samples but were noncoding in other samples. We defined these transcripts as CDCTs (context-dependent coding transcripts). In further model building, we retained genes with all isoforms as coding and selected one representative isoform (expressed in the most samples) for the gene into the training data (Supplementary Figure 5).

Collection of classified features based on biological knowledge

Based on biological knowledge, the candidate features were collected from manual literature survey. Our aim is to depict translation control *in vivo*, so the first criterion for feature collection is able to be explained by translation-related biological process. The second criterion is the feature value should be easily encoded by sequence and transcription information. As a result, we collected 221 features and grouped them into 5 translation-related processes. 1) Expression abundance: in addition to RNA-seq expression level, miRNA targeting also affect RNA abundance. Thus, we scanned miRNA target sites on 3'UTR and incorporated it into a feature. 2) Translation initiation: the RNA folding energy and sequence context around the translation initiation site are related to the transcripts' coding ability, so we used these features in RiboCalc. 3) Translation elongation: the translation elongation rate is often altered by the adaption between codon usage and the corresponding tRNA abundance[10]. Therefore, the frequency of 64 codons, as well as some other indexes describing codon usage bias, were used in this class of features. 4) Translation regulators: the abundance of translation regulatory factors are related to translation level of RNAs[31], thus, expression level of translation-related genes annotated by GO were taken into account as features. 5) Transcript structure: other sequence features are also related to transcript coding ability, such as length and UTR GC content. The mechanism of these features might not be clearly validated with experiments but we added them into the list.

Considering the genomic mutations and RNA editing, we modified transcript reference sequences with variations called by GATK[32] in RNA-seq, and calculated feature value based on modified sequences. The mutations causing start codon loss were ignored (see "Mutated transcript sequence identification with RNA-seq data" at https://github.com/gao-lab/RiboCalc/blob/master/feature_calculation/RNAandRibo-seq_processing.txt). The ORFs that we used for feature calculation are described in

Supplementary Figure 3. All the features and the calculation method are shown in Supplementary Table 4. We provide feature calculation scripts at https://github.com/gao-lab/RiboCalc/tree/master/feature_calculation.

Building cell-specific models

We chose 5 cell lines for cell-specific model building. The 5 cell lines were the most common ones among our collected resources (Supplementary Table 1) and could represent 5 different tissues. We used Ribo-based coding transcripts for model building and removed redundant sequences with more than 90% identity using CD-hit[33]. We randomly selected 3,000 transcripts as training data and the rest as testing data for each model. The models with selected features were built through linear regression with the Lasso penalty. The feature data and model building script are at https://github.com/gao-lab/RiboCalc/tree/master/cell_specific_model.

RiboCalc model building

By adding *trans*- features of translation regulators, we built an "environment-aware" model for quantitative prediction of coding ability globally and named the model RiboCalc. The detailed methods of model building were as follows:

We pooled coding transcripts from all cell lines together. If identical transcripts expressed in more than one ribosome profiling sample, we selected the one with median TPM in the corresponding RNA-seq data. The dataset consists of 8,193 transcripts, and we randomly split them into 5,000 training cases and 3,193 testing cases.

All the values of each feature were scaled to the interval of [0, 1] for training data as following equation (min-max normalization).

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Since gene-level expression abundance estimation was reported to be more accurate than isoform level [34-36], we took gene TPM for calculation. Since different studies had a various distribution of expression level in RNA-seq and Ribo-seq, to pool all the data together, we adopted cross-sample normalization for RNA-TPM and Ribo-TPM. The RNA-TPM and Ribo-TPM were normalized with TPM of the housekeeping gene *HPRT1* using the following equation. *HPRT1* was selected based on the work of Valente *et al.*[37].

$$\mathrm{RiboTPM}'_{i,j} = \mathrm{RiboTPM}_{i,j} * \frac{\mathrm{median}(\mathrm{RiboTPM}_{HPRT1})}{\mathrm{RiboTPM}_{HPRT1,j}}$$

The Ribo-TPM of transcript *i* in sample *j* was scaled by the ratio of the *HPRT1* Ribo-TPM in sample *j* with the median *HPRT1* Ribo-TPM among all the samples. The same normalization strategy was also applied to RNA-TPM.

We first removed highly correlated features (Pearson's *r* above 0.9) with "findCorrelation" in the R caret package[38]. Then, feature selection was implemented through a linear model with Lasso regularization. We searched the parameter λ with the minimum mean squared error (MSE) in 5-fold cross validation. The Lasso regression was implemented by the glmnet package[39] in R. We provide raw data and script at https://github.com/gao-lab/RiboCalc/tree/master/RiboCalc.

Human model comparison

OCTOPOS [17]. We download the raw data OCTOPOS used for HEK 293. The RNA-seq data was from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE38356. Protein abundance data was downloaded from https://pax-db.org/dataset/9606/329/. Since correlation calculation would be affected by data size, we randomly selected same amount of transcripts (461) with OCTOPOS as testing set. We provide the script of feature calculation and model testing from OCTOPOS data at

https://github.com/gao-lab/RiboCalc/blob/master/feature_calculation/script/test_OCTOPOS.sh. Users could follow the script to apply RiboCalc on their own data.

Li's human model [40]. We downloaded their data "Additional file 2" at https://doi.org/10.1186/s13059-019-1761-9. The UTR and CDS sequences were obtained from Ensembl Genes 104 with the transcript IDs. Since Li's human model used RPKM as expression level which was hardly transformed into TPM without knowing the full transcriptome[41], we retrained RiboCalc with their data. We randomly selected 2,000 transcripts as testing data and the rest as training data. The generated feature data and testing script are at https://github.com/gao-lab/RiboCalc/tree/master/human_model_comparison/LiJJ.

Sample's model [42]. We downloaded the model from https://github.com/pjsample/human_5utr_modeling/tree/master/modeling/saved_models/main_MRL_model.hdf5. Given that Sample *et al.*'s model requires the input sequence length to be 50, we generated 50nt fragments (window size = 50, step size = 1) of 5'UTRs of RiboCalc testing data. The 5'UTR sequences were downloaded from Ensembl Genes 104 and the transcripts without 5'UTR annotation were removed in the testing set. We used the average predicted value of all 50nt windows from 5'UTR sequences as the final predicted value of the transcripts. See https://github.com/gao-lab/RiboCalc/tree/master/human_model_comparison/SamplePJ.

Comparison with Li's model in yeast

Li *et al*. used transcript sequence features to predict translation rate (TR) in yeast [6]. In their definition, TR is defined as the number of protein molecules translated per mRNA molecule, which is the ratio of ribosome density (also abbreviated to Ribo-TPM for uniformity with RiboCalc) to RNA expression abundance in Ribo-seq.

Thus, RiboCalc could predict TR by dividing RNA abundance into the output as the equation below.

$$\mathrm{RiboTPM} = \mathrm{RNA} * \mathrm{TR}$$

The yeast transcript sequences, ORFs, expression abundance and ribosome density were obtained from the "nar-00812-a-2017-File019.csv" file in the supplementary data downloaded from https://doi.org/10.1093/nar/gkx898. Since the RiboCalc prediction depends on the 3'UTR sequence which is not provided in Li's data, we fetched the 3'UTR sequences from the Saccharomyces Genome Database (SGD) [43] and removed 909 transcripts without 3'UTR annotation.

To build the RiboCalc yeast model, we randomly split the remaining 1,541 transcripts into 1,000 training cases and 541 testing cases. After considering the systematic differences between yeast and human models, we removed several human-specific features from the yeast model (Supplementary Table 5). The yeast model training approach was identical to the RiboCalc human model. To make a fair comparison, we retrained Li's yeast model using the new training set by strictly following their description. The performance of Li's retrained model was similar to their original report (R^2 of TR are both 0.80). The correlations in Table 3 were calculated from the testing set. See https://github.com/gao-lab/RiboCalc/tree/master/RiboCalc_yeast for raw data and script.

Ribo-lncRNA analysis

Long noncoding RNAs (lncRNAs) associated with ribosomes are abbreviated to ribo-lncRNAs here. The ribo-lncRNAs from Ruiz-Orera *et al.* were identified from the GSE22004 dataset in the NCBI GEO database [20]. Thus, we used GSM546926 as the Ribo-seq sample and GSM546927 as the RNA-seq sample in GSE22004 for RiboCalc analysis. The "top coding score lncRNAs", "lncRNAs with homologies"

and “young codRNAs” were identified by Ruiz-Orera *et al.* and downloaded from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4359382/bin/elife03523s002.xls. The “non-ribo-lincRNAs” and “ribo-lincRNAs” and “other codRNAs” were from our analysis of the data. “Non-ribo-lincRNAs” refers to long intergenic noncoding RNAs (lincRNAs) without ribo-reads, while “Ribo-lincRNAs” are the lincRNAs covered by ribo-reads. The “other codRNAs” are protein-coding transcripts with translated ORFs in ribosome profiling excluding the young codRNAs from the original report. In this study, transcripts with FPKM lower than 0.2 were excluded.

For ribo-lncRNAs identified by Zeng *et al.* [44], their resources were included in the Ribo-seq samples collected by us, except for the data associated with drug treatment. Therefore, we directly retrieved the features of those ribo-lncRNAs in our data. The four classes of ribo-lncRNAs in Figure 3C were downloaded from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5975437/bin/12864_2018_4765_MOESM10_ESM.xlsx. See https://github.com/gao-lab/RiboCalc/tree/master/ribo-lncRNA for raw data and script.

Results

Transcripts' coding status identification in ribosome profiling data

To accurately quantify the coding ability of transcripts in various cells, we retrieved 61 pairs of reliable Ribo-seq data coupled with RNA-seq data from the NCBI GEO database [19], covering 1 tissue and 21 cell lines (Supplementary Table 1). The expression abundance in the corresponding Ribo-seq data (abbreviated to Ribo-TPM) were employed as the quantitative metric for the coding ability. For accurate prediction of coding ability, the first step is to obtain protein-coding transcripts for model building. Thus, by applying rigorous filtering criteria, we called translation status for 101,170 out of 199,169 GENCODE gene models (the rest were filtered out due to either a lack of expression signals in the chosen samples (96,968) or a failure of calling translated ORF reliably (1,031), see Supplementary Figure 5 for detailed procedure). Among the 101,170 transcripts, the translation status of 46% were found to be "coding" while 43% were "noncoding" in all samples (Figure 1B). Interestingly, we also found that 11% of the transcripts exhibited diverse coding ability among cell lines (i.e., coding in some cell lines but noncoding in others), and we named them context-dependent coding transcripts (CDCTs).

[Insert Figure 1 here]

Figure 1 Transcript coding status classification in ribosome profiling data. A) Workflow of RiboCalc model building **B)** The percentage of transcripts with a particular coding ability classification. The blue fraction shows noncoding transcripts, the yellow fraction shows coding transcripts, and the green fraction shows CDCTs. **C)** Comparison between coding ability classification based on ribosome profiling with the biotypes annotated by GENCODE. The left bar shows the transcripts annotated as coding transcripts by GENCODE, while the right bar shows noncoding transcripts. The fractions of each bar correspond to the Ribo-seq based coding status calling in our analysis, and the colors of particular types are the same as those in Figure 1B.

RiboCalc: predicting transcript coding ability in human

To identify features contributing to coding ability determination, we first compiled a candidate list of intrinsic ("*cis*-") and contextual ("*trans*-") features after systematic literature survey (Supplementary Table 4, also see "**Collection of classified features based on biological knowledge**" in "**Materials and Methods**"). The intrinsic features represent transcript sequence characteristics. And we grouped them into three categories: "translation initiation"[11], "translation elongation"[45] and "transcript structure" [6, 13] based on the underlying biological process. The contextual features were collected to depict the environment for translation in the cell and all belong to the category "expression abundance" [46]). We then incorporated these features together to build cell-specific models for five representative cell lines with a Lasso regression based feature selection. All models works well (Table 1 and Supplementary Figure 6), highlighting the effectiveness of these selected features (Supplementary Table 4).

[Insert Table 1 here]

Table 1 Ribo-seq resources and transcript dataset size for cell-specific model building

The 5 cell-specific models showed a significant correlation between the predicted values and observed Ribo-TPM on the testing set. These cell lines selected were the five that covered most studies in all of our collected samples (Supplementary Table 1). The data size were the number of coding transcripts identified following the stringent criteria (see "**Identification of translated ORFs**" and "**Calling transcript's coding status**" in "**Materials and Methods**"). As reported in Table 1, the lowest *r* of the 5 models was 0.78. The remaining 4 models all had a correlation above 0.8, among which the HeLa model had the highest correlation of approximately 0.89.

We further tried to incorporate the effect of cell "environment" by introducing expression level of *trans*-factors (i.e., translation regulators) in the corresponding sample ("Translation regulators" in Supplementary Table 4). The "environment-aware"

model (RiboCalc) accurately predicts coding ability (Ribo-TPM) globally, with r = 0.81 in in testing data from all 22 cell lines (Figure 2A). The performance of RiboCalc is comparable with 5 cell-specific models (Supplementary Figure 6), suggesting its prediction efficiency across cells by adding "environment" features.

We compared RiboCalc's performance with Sample *et al.*'s model [42] and Li *et al.*'s model [40] in human. Li's human model predicted TR (translation rate) which could be calculated as the ratio between Ribo-seq and RNA-seq abundance. RiboCalc predicted TR with Pearson's r = 0.66, which is higher than Li's human model (r = 0.64, Supplementary Table 6). Sample et al. built a deep learning model to predict ribosome loading. Since Sample *et al.* only provided model data of 50nt 5'UTR sequences which are not sufficient for RiboCalc prediction, we applied Sample's model directly on RiboCalc testing data. It showed a much lower correlation (r = 0.18, Supplementary Figure 7) than RiboCalc (r = 0.81, Figure 2A). Therefore, RiboCalc accurately predicts ribosome density with intrinsic ("*cis*-") and contextual ("*trans*-") features.

Parameters of RiboCalc are interpretable by biological impact on translation

As RiboCalc is a linear model fitted by normalized feature values, features' coefficients quantify their contribution (e.g. positive coefficients suggest facilitation of coding ability, while negative coefficients suggest an adverse effect). A manually checking in literatures effectively connects model parameters with prior biological knowledge (Table 2). For instance, the feature with the highest positive coefficient is the RNA abundance (i.e. TPM), confirming existing reports on the dominant influence of transcript expression level on protein translation [46, 47]. Similarly, being consistent with the observation that longer transcripts reduced the number of dropped ribosomes diffusing to the translation initiation site as well as mRNA circularization[13], the model also demonstrates significant adverse effect for the

length of transcript.

[Insert Table 2 here]

Table 2 Feature coefficient with greatest contribution in RiboCalc and the effect on translation in the literature

The table shows features with the most positive or negative coefficient in RiboCalc model. Since all features were scaled into the interval of [0, 1], the absolute value and sign symbol of their coefficients could be interpreted as the impact on coding ability.

Based on biological knowledge, we grouped all features into 5 translation-related processes, as translation initiation, translation elongation and transcript structure for intrinsic features from sequence, and expression abundance and translation regulators as contextual features for environment. By calculating Ribo-TPM through a single class of features with RiboCalc, we compared their performance with the correlation between predicted and observed value. The predicted results of both intrinsic and contextual features showed a significant correlation with the observed Ribo-TPM (Figure 2B). And consistent with previous studies [46], expression abundance showed the greatest importance for coding ability in RiboCalc (Figure 2C).

To further validate the model's effectiveness, we applied it to unicellular organism yeast based on published dataset[6], and found that the original RiboCalc model accurately predicted both coding ability and translation rate, with comparable performance to the state-of-arts model in yeast (Table 3, see "**Comparison with Li's model in yeast**" section in "**Materials and Methods**" for details). Given the systematic differences between human and yeast, we also retrained a RiboCalc yeast model, with yeast data as input but adopting exactly the same feature set and fitting procedure as in the human model. As expected, the RiboCalc yeast model further improved performance overall (Table 3).

[Insert Table 3 here]

Table 3 Correlation between predicted and observed values in RiboCalc human, RiboCalc yeast and Li's model

Intriguingly, we found that, through the same set of model features adopted, multiple coefficients of the RiboCalc yeast model differ from those of the human model. The feature coefficient with the largest difference is the length of 3'UTR (Figure 2D). Recently, Fu *et al.* demonstrated that a longer 3'UTR increased the possibility of miRNA targeting in mammalian cells, resulting in a reduction in protein translation[12], whereas no clear evidence for the pervasive existence of miRNAs in yeast. Consistently, the coefficient for the 3'UTR length in the human model is a large negative value and close to zero in yeast (Supplementary Table 7, also see Figure 2D and 2E for another case on codon usage). We believe that divergent pattern reflected by RiboCalc models could facilitate investigating the inter-species discrepancy of translation regulatory mechanisms.

[Insert Figure 2 here]

Figure 2 Model performance and feature contribution of RiboCalc. A) Scatter plot of RiboCalc predicted and observed values. The x axis shows the observed Ribo-TPM of coding transcripts in the testing set, while the y axis shows the corresponding Ribo-TPM predicted by RiboCalc. Pearson's *r* and the significance level were calculated between x and y scores. **B)** Effectiveness of intrinsic and contextual features in RiboCalc. The bars show Pearson's *r* between the Ribo-TPM predicted by each single class of features and the observed value. The p-value above the bars indicates the significance of the correlation. **C)** Feature importance of 5 processes in RiboCalc. The boxes show the feature coefficient distributions of the 5 translation-related processes. Since all the feature values were scaled to [0, 1], the feature coefficients could be taken as contributions to coding ability prediction in this plot. The numbers above the boxes are the corresponding feature numbers of each class. **D)** Scatter plot of feature coefficients in RiboCalc human and yeast models. The x axis shows the feature coefficients in the RiboCalc human model.

The y axis shows the corresponding feature coefficients in the RiboCalc yeast model. The colors of the points represent the four quadrants. The black text labels the feature points of transcript length, 3'UTR length and RNA TPM. **E)** The feature names and coefficient values of RiboCalc human and yeast models in each quadrant. The text on the labels stand for the feature names (Supplementary Table 4). The codon name stands for codon usage frequency, "init_fold" means translation initiation sequences' minimum free energy (MFE) predicted by RNAfold. The points of transcript length and RNA TPM are not shown in this plot due to the limitations of the axes. The x axis, y axis and colors are the same as those in Figure 2D.

Discussion

To model transcript coding ability quantitatively, we proposed and implemented RiboCalc. RiboCalc not only effectively predicts coding ability but also reveals several intriguing novel hints for deconvoluting the translation control mechanism. For example, according to the model, the GC content at 3' UTR presents a positive contribution to the coding ability. Coupling with recent report on AU-rich 3'UTRs leads to decreased stability of RNAs[48], we could reasonably infer that the 3'UTR with higher GC content would promote translation efficiency by slowing RNA decay. Meanwhile, we also notice a few inconsistencies between model-estimated coefficients and existing literature. For example, a reported translation-suppressing gene, *PAIP2B*, showed a positive contribution to coding ability in our model, with statistically significant positive correlation detected for its abundance and the median Ribo-TPM of translatable genes in the corresponding sample (Supplementary Figure 8, also see Supplementary Figure 9). All these observations shed lights on further mechanism study and validation.

These insights could lead to several potential applications like improving existing codon optimization tool [17, 49]. A direct comparison with OCTOPOS, a recent published mechanism-oriented codon optimization tool [17], found that the RiboCalc model could effectively predict protein abundance (r = 0.63, HEK293 human dataset, vs. r = 0.61 reported in the original paper, Supplementary Figure 10A), even given the fact that the RiboCalc model was trained to predict Ribo-seq level, instead of protein abundance measured in MS as what the OCTOPOS did. Of interest, when being retrained with OCTOPOS HEK293 dataset, the RiboCalc model showed improved prediction accuracy (r = 0.73, Supplementary Figure 10B, also see "**Human model comparison**" in "**Materials and Methods**"), highlighting RiboCalc's potential to pinpoint novel translation-regulation-related features.

The RiboCalc model suggests that both intrinsic ("*cis*-", like the transcript sequence) and contextual ("*trans*-", like expression level of transcript and translation regulators) features contribute to the transcript coding ability decision in cells. Intriguingly, we found a number of transcripts exhibited diverse coding ability among cell lines (CDCTs, Supplementary Figure 11, also see detailed list at https://github.com/gao-lab/RiboCalc/blob/master/CDCT/CDCT_transcript_list.tab). Among the protein-coding transcripts annotated by GENCODE, 15% were classified as noncoding and 17% as CDCTs; meanwhile, one-third of GENCODE-annotated ncRNAs were classified as coding, and 7% as CDCTs (Figure 1C, also see the list at https://github.com/gao-lab/RiboCalc/blob/master/CDCT/CDCT_lincRNA_TransLnc_evidence_number.tab). The model shows that CDCTs under coding context get higher RiboCalc scores than these under noncoding context (Figure 3A). Consistently, canonical protein coding genes have been validated experimentally that protein expression ability could be varied or even silenced without altering mRNA transcribing [50-52] and several annotated lncRNAs were found to be associated with ribosomes (abbreviated to ribo-lncRNAs) [20, 44, 53] and encode functional peptides [54-56]. RiboCalc model confirms that reported ribo-lncRNAs as well as lncRNAs with high homologies and top coding score[20, 44] have coding ability significantly higher than that of non-ribo-lncRNAs, and close to those of young experimentally validated coding RNAs (Figure 3B and 3C, also see Supplementary Figure 12 for more analysis on GENCODE transcripts). Collectively, these results suggest that transcript's coding ability should be modeled as a context-dependent continuous value, rather than a certain binary class.

[Insert Figure 3 here]

Figure 3 Biological interpretation of transcripts with ambiguous coding ability from the RiboCalc model. A) RiboCalc prediction of transcripts with a particular coding

ability classification. “Noncoding” refers to the noncoding transcripts identified in Ribo-seq, “coding” is the testing data of RiboCalc. The “coding CDCTs” are CDCTs present under coding context (observed as coding in particular samples), and “noncoding CDCTs” are under noncoding context. **B)** Prediction of ribo-lncRNAs from Ruiz-Orera *et al.* in RiboCalc. The boxes show the predicted Ribo-TPM distribution of a particular class of RNAs. The “non-ribo-lncRNA” are lincRNAs without ribosome coverage, while “ribo-lncRNA” are covered by ribo-reads. The “top coding score lncRNA” are lncRNAs with the highest sequence similarity with protein-coding transcripts. The “lncRNA with homologies” are lncRNAs conserved among species. “Young codRNA” are validated coding RNAs with a short evolutionary history, while “other codRNA” are the rest coding RNAs. **C)** Prediction of ribo-lncRNAs from Zeng *et al.* in RiboCalc. The boxes show the predicted Ribo-TPM distribution of a particular class of RNAs. The “trans-lncRNA” are translated lncRNAs, “ribo-lncRNA” are lncRNAs only covered by ribo-reads, “non-ribo-lncRNA” are lncRNAs without ribo-reads, and “other” refer to unexpressed lncRNAs. All significance levels in Figure 3A, B and C are based on Wilcox test.

Data Availability

All scripts and data are available online at https://github.com/gao-lab/RiboCalc/.

Acknowledgments

The authors thank Drs. Zemin Zhang, Cheng Li, Letian Tao, Jian Lu and Liping Wei at Peking University for their helpful comments and suggestions during the study. Part of the analysis was performed on the Computing Platform of the Center for Life Sciences of Peking University and supported by the High-performance Computing Platform of Peking University.

Key Points

- We built an *in silico* model for predicting transcripts' coding ability accurately in human.
- We showed, quantitatively, that both intrinsic and contextual features contribute to coding ability determination.
- We identified a great number of transcripts are with distinct coding abilities among different type of cells (i.e. context-dependent coding transcripts, CDCTs), suggesting the transcript’s coding ability should be modeled as a context-dependent continuous spectrum, rather than a static binary classification as “coding” or “noncoding”.

Funding

This work was supported by funds from the National Key Research and Development Program (2016YFC0901603), the China 863 Program (2015AA020108), as well as

the State Key Laboratory of Protein and Plant Gene Research and the Beijing Advanced Innovation Center for Genomics (ICG) at Peking University. The research of G.G. was supported in part by the National Program for Support of Top-notch Young Professionals.

Conflict of interest statement. None declared.

References

1. Ashe MP, De Long SK, Sachs AB. Glucose depletion rapidly inhibits translation initiation in yeast, Mol Biol Cell 2000;11:833-848.
2. Bazin J, Baerenfaller K, Gosai SJ et al. Global analysis of ribosome-associated noncoding RNAs unveils new modes of translational regulation, Proc Natl Acad Sci U S A 2017;114:E10018-E10027.
3. Liu S, Hausmann S, Carlson SM et al. METTL13 Methylation of eEF1A Increases Translational Output to Promote Tumorigenesis, Cell 2019;176:491-504 e421.
4. Huang T, Wan S, Xu Z et al. Analysis and prediction of translation rate based on sequence and functional features of the mRNA, PLoS One 2011;6:e16036.
5. Dvir S, Velten L, Sharon E et al. Deciphering the rules by which 5'-UTR sequences affect protein expression in yeast, Proc Natl Acad Sci U S A 2013;110:E2792-2801.
6. Li JJ, Chew GL, Biggin MD. Quantitating translational control: mRNA abundance-dependent and independent contributions and the mRNA sequences that specify them, Nucleic Acids Res 2017;45:11821-11836.
7. Zur H, Tuller T. Transcript features alone enable accurate prediction and understanding of gene expression in S. cerevisiae, BMC Bioinformatics 2013;14 Suppl 15:S1.
8. Andreev DE, Dmitriev SE, Loughran G et al. Translation control of mRNAs encoding mammalian translation initiation factors, Gene 2018;651:174-182.
9. Gingold H, Pilpel Y. Determinants of translation efficiency and accuracy, Mol Syst Biol 2011;7:481.
10. Dana A, Tuller T. The effect of tRNA levels on decoding times of mRNA codons, Nucleic Acids Res 2014;42:9171-9181.
11. Zhang S, Hu H, Jiang T et al. TITER: predicting translation initiation sites by deep learning, Bioinformatics 2017;33:i234-i242.
12. Fu Y, Chen L, Chen C et al. Crosstalk between alternative polyadenylation and

miRNAs in the regulation of protein translational efficiency, Genome Res 2018;28:1656-1663.
13. Fernandes LD, Moura APS, Ciandrini L. Gene length as a regulator for ribosome recruitment and protein synthesis: theoretical insights, Sci Rep 2017;7:17409.
14. Lander ES, Linton LM, Birren B et al. Initial sequencing and analysis of the human genome, Nature 2001;409:860-921.
15. Tamarkin-Ben-Harush A, Schechtman E, Dikstein R. Co-occurrence of transcription and translation gene regulatory features underlies coordinated mRNA and protein synthesis, BMC Genomics 2014;15:688.
16. Volkova OA, Kondrakhin YV, Yevshin IS et al. Assessment of translational importance of mammalian mRNA sequence features based on Ribo-Seq and mRNA-Seq data, J Bioinform Comput Biol 2016;14:1641006.
17. Trosemeier JH, Rudorf S, Loessner H et al. Optimizing the dynamics of protein expression, Sci Rep 2019;9:7511.
18. Xie SQ, Nie P, Wang Y et al. RPFdb: a database for genome wide information of translated mRNA generated from ribosome profiling, Nucleic Acids Res 2016;44:D254-258.
19. Wheeler DL, Barrett T, Benson DA et al. Database resources of the National Center for Biotechnology Information, Nucleic Acids Res 2005;33:D39-45.
20. Ruiz-Orera J, Messeguer X, Subirana JA et al. Long non-coding RNAs as a source of new peptides, Elife 2014;3:e03523.
21. Pertea M, Shumate A, Pertea G et al. CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise, Genome Biol 2018;19:208.
22. van Heesch S, Witte F, Schneider-Lunitz V et al. The Translational Landscape of the Human Heart, Cell 2019;178:242-260 e229.
23. Erhard F, Halenius A, Zimmermann C et al. Improved Ribo-seq enables identification of cryptic translation events, Nat Methods 2018;15:363-366.
24. Vizcaino JA, Csordas A, del-Toro N et al. 2016 update of the PRIDE database and its related tools, Nucleic Acids Res 2016;44:D447-456.
25. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics 2006;22:1658-1659.
26. Chi H, Liu C, Yang H et al. Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine, Nat Biotechnol 2018.
27. Xiao Z, Huang R, Xing X et al. De novo annotation and characterization of the translatome with ribosome profiling data, Nucleic Acids Res 2018;46:e61.
28. Ji Z. RibORF: Identifying Genome-Wide Translated Open Reading Frames Using Ribosome Profiling, Curr Protoc Mol Biol 2018;124:e67.
29. Baboo S, Cook PR. "Dark matter" worlds of unstable RNA and protein, Nucleus 2014;5:281-286.
30. Ning K, Fermin D, Nesvizhskii AI. Comparative analysis of different label-free mass spectrometry based protein abundance estimates and their correlation with

RNA-Seq gene expression data, J Proteome Res 2012;11:2261-2271.
31. Holcik M, Sonenberg N. Translational control in stress and apoptosis, Nat Rev Mol Cell Biol 2005;6:318-327.
32. McKenna A, Hanna M, Banks E et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, Genome Res 2010;20:1297-1303.
33. Fu L, Niu B, Zhu Z et al. CD-HIT: accelerated for clustering the next-generation sequencing data, Bioinformatics 2012;28:3150-3152.
34. Kanitz A, Gypas F, Gruber AJ et al. Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data, Genome Biol 2015;16:150.
35. Sun J, Chang JW, Zhang T et al. Platform-integrated mRNA isoform quantification, Bioinformatics 2020;36:2466-2473.
36. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences, F1000Res 2015;4:1521.
37. Valente V, Teixeira SA, Neder L et al. Selection of suitable housekeeping genes for expression analysis in glioblastoma using quantitative RT-PCR, BMC Mol Biol 2009;10:17.
38. Max K, Contributions from Jed W, Steve W et al. caret: Classification and Regression Training 2016.
39. Tibshirani JFaTHaR. Regularization Paths for Generalized Linear Models via Coordinate Descent, JOURNAL OF STATISTICAL SOFTWARE 2010;33:1--22.
40. Li JJ, Chew GL, Biggin MD. Quantitative principles of cis-translational control by general mRNA sequence features in eukaryotes, Genome Biol 2019;20:162.
41. Wagner GP, Kin K, Lynch VJ. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples, Theory Biosci 2012;131:281-285.
42. Sample PJ, Wang B, Reid DW et al. Human 5' UTR design and variant effect prediction from a massively parallel translation assay, Nat Biotechnol 2019;37:803-809.
43. Cherry JM, Hong EL, Amundsen C et al. Saccharomyces Genome Database: the genomics resource of budding yeast, Nucleic Acids Res 2012;40:D700-705.
44. Zeng C, Fukunaga T, Hamada M. Identification and analysis of ribosome-associated lncRNAs using ribosome profiling data, BMC Genomics 2018;19:414.
45. Zur H, Tuller T. RFMapp: ribosome flow model application, Bioinformatics 2012;28:1663-1664.
46. Csardi G, Franks A, Choi DS et al. Accounting for experimental noise reveals that mRNA levels, amplified by post-transcriptional processes, largely determine steady-state protein levels in yeast, PLoS Genet 2015;11:e1005206.
47. Franks A, Airoldi E, Slavov N. Post-transcriptional regulation across human tissues, PLoS Comput Biol 2017;13:e1005535.

48. Guhaniyogi J, Brewer G. Regulation of mRNA stability in mammalian cells, Gene 2001;265:11-23.
49. Dana A, Tuller T. Mean of the typical decoding rates: a new translation efficiency index based on the analysis of ribosome profiling data, G3 (Bethesda) 2014;5:73-80.
50. Pelletier J, Graff J, Ruggero D et al. Targeting the eIF4F translation initiation complex: a critical nexus for cancer development, Cancer Res 2015;75:250-263.
51. Robert F, Pelletier J. Exploring the Impact of Single-Nucleotide Polymorphisms on Translation, Front Genet 2018;9:507.
52. Djuranovic S, Nahvi A, Green R. miRNA-mediated gene silencing by translational repression followed by mRNA deadenylation and decay, Science 2012;336:237-240.
53. Wang H, Wang Y, Xie S et al. Global and cell-type specific properties of lincRNAs with ribosome occupancy, Nucleic Acids Res 2017;45:2786-2796.
54. Matsumoto A, Pasut A, Matsumoto M et al. mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide, Nature 2017;541:228-232.
55. Pauli A, Norris ML, Valen E et al. Toddler: an embryonic signal that promotes cell movement via Apelin receptors, Science 2014;343:1248636.
56. Pang Y, Mao C, Liu S. Encoding activities of non-coding RNAs, Theranostics 2018;8:2496-2507.
57. Lee AS, Kranzusch PJ, Cate JH. eIF3 targets cell-proliferation messenger RNAs for translational activation or repression, Nature 2015;522:111-114.
58. Tao X, Gao G. Tristetraprolin Recruits Eukaryotic Initiation Factor 4E2 To Repress Translation of AU-Rich Element-Containing mRNAs, Molecular and Cellular Biology 2015;35:3921-3932.
59. Fonseca BD, Zakaria C, Jia JJ et al. La-related Protein 1 (LARP1) Represses Terminal Oligopyrimidine (TOP) mRNA Translation Downstream of mTOR Complex 1 (mTORC1), J Biol Chem 2015;290:15996-16020.
60. Fu YG, Chen LT, Chen CY et al. Crosstalk between alternative polyadenylation and miRNAs in the regulation of protein translational efficiency, Genome Research 2018;28:1656-1663.

Tables

Table 1 Ribo-seq resources and transcript dataset size for cell-specific model building

Sample ID	Cell line	Bioresources	Data size	Pearson's *r* with Ribo-TPM
SRR1803151	GM12891	B lymphocyte	4,781	0.785
SRX870805	HEK 293	Embryonic kidney cell	6,057	0.836
SRR970565	HeLa	Cervical cancer cell	6,440	0.886
SRR627625	BJ	Foreskin fibroblast	5,746	0.830
SRR3208870	hESC.2	Embryonic stem cell	6,260	0.862

The 5 cell-specific models showed a significant correlation between the predicted values and observed Ribo-TPM on the testing set. These cell lines selected were the five that covered most studies in all of our collected samples (Supplementary Table 1). The data size were the number of coding transcripts identified following the stringent criteria (see "**Identification of translated ORFs**" and "**Calling transcript's coding status**" in "**Materials and Methods**"). As reported in Table 1, the lowest *r* of the 5 models was 0.78. The remaining 4 models all had a correlation above 0.8, among which the HeLa model had the highest correlation of approximately 0.89.

Table 2 Feature coefficient with greatest contribution in RiboCalc and the effect on translation in the literature

Top 15 positive feature	Coefficient	Supported by literature	Evidence
RNA_TPM	0.623	yes	[46, 47]
EIF3L	0.314	yes	https://www.genecards.org/cgi-bin/carddisp.pl?gene=EIF3L
EIF2B3	0.146	yes	https://www.genecards.org/cgi-bin/carddisp.pl?gene=EIF2B3
RPS9	0.117	yes	https://www.uniprot.org/uniprot/P46781
PAIP2B	0.117	no	https://www.uniprot.org/uniprot/Q9ULR5
AAG	0.097	yes	https://www.cs.tau.ac.il/~tamirtul/MTDR/mu_vals.html
ATG	0.096	ambivalent	https://www.cs.tau.ac.il/~tamirtul/MTDR/mu_vals.html
MTDR	0.096	yes	[49]
RARA	0.092	no	https://www.uniprot.org/uniprot/P10276
EEF1G	0.090	yes	https://www.genecards.org/cgi-bin/carddisp.pl?gene=EEF1G
GAT	0.088	ambivalent	https://www.cs.tau.ac.il/~tamirtul/MTDR/mu_vals.html
C12orf65	0.078	yes	https://ghr.nlm.nih.gov/gene/C12orf65
RPS14	0.078	yes	https://www.uniprot.org/uniprot/P62263
3UTR_GC	0.077	yes	[48]
MTIF2	0.075	yes	https://www.uniprot.org/uniprot/P46199
Top 15 negative	**Coefficient**	**Supported by**	**Evidence**

feature		literature	
EIF3G	-0.315	ambivalent	[57]
Length	-0.184	yes	[13]
MTRF1	-0.152	yes	https://www.genecards.org/cgi-bin/carddisp.pl?gene=MTRF1
EIF2A	-0.151	ambivalent	https://www.uniprot.org/uniprot/P26641
TYMS	-0.116	yes	https://www.uniprot.org/uniprot/P04818
AGC	-0.104	no	https://www.cs.tau.ac.il/~tamirtul/MTDR/mu_vals.html
EIF4E2	-0.098	yes	https://string-db.org/network/9606.ENSP00000258416 [58]
EIF4E	-0.093	yes	https://www.uniprot.org/uniprot/P06730
LARP1	-0.092	yes	https://www.uniprot.org/uniprot/Q6PKG0 [59]
EEF1D	-0.092	ambivalent	https://www.uniprot.org/uniprot/P29692
EIF3E	-0.088	ambivalent	https://www.uniprot.org/uniprot/P60228
3UTR_length	-0.088	yes	[60]
MTRF1L	-0.088	yes	https://www.genecards.org/cgi-bin/carddisp.pl?gene=MTRF1L
CGG	-0.084	no	https://www.cs.tau.ac.il/~tamirtul/MTDR/mu_vals.html
EIF2AK3	-0.081	yes	https://www.uniprot.org/uniprot/Q9NZJ5

The table shows features with the most positive or negative coefficient in RiboCalc model. Since all features were scaled into the interval of [0, 1], the absolute value and sign symbol of their coefficients could be interpreted as the impact on coding ability.

Table 3 Correlation between predicted and observed values in RiboCalc human, RiboCalc yeast and Li's model

Predicted value	**Model**	**Pearson's *r***	**Spearman's *r***
TR	RiboCalc human	0.759	0.762
	RiboCalc yeast	0.886	0.898
	Li yeast	0.874	0.887
Ribo-TPM	RiboCalc human	0.974	0.969
	RiboCalc yeast	0.987	0.984
	Li yeast	0.986	0.982