RIBO-former: applying attention-based neural networks for precise delineation of translated open reading frames using ribosome profiling data.

Jim Clauwaert*
Michigan Medicine
Pediatric Hematology/Oncology,
University of Michigan
United States

Zahra McVey
Novo Nordisk Research
Centre Oxford,
Novo Nordisk Ltd.
Oxford, United Kingdom

Ramneek Gupta
Novo Nordisk Research
Centre Oxford,
Novo Nordisk Ltd.
Oxford, United Kingdom

John Prensner
Michigan Medicine
Pediatric Hematology/Oncology,
University of Michigan
United States

Gerben Menschaert
Department of Data Analysis and
Mathematical Modelling,
Ghent University, Belgium

Abstract

Ribosome profiling is a next-generation sequencing technique used to chart translation by means of mRNA ribosome occupancy. It is an instrumental tool for the detection of non-canonical coding sequences. However, due to the complex nature of next-generation sequencing data, existing solutions that seek to identify translated open reading frames from the data still suffer from a high degree of false positive predictions. We propose RIBO-former, a new approach featuring several innovations for the detection of translated open reading frames. RIBO-former is built using recent transformer models that have achieved considerable advancements in the field of natural language processing. We discuss several strategies to parse ribosome profiling data for which we provide a comprehensive ablation study. Through benchmarking, we find RIBO-former to outperform previous methodologies by a large margin.

Introduction

Ribosome profiling was first introduced by Ingolia et al. to study the translatome of cells through deep sequencing (2009). The technique sequences ribosome-protected fragments that are aligned to a reference genome, resulting in an occupancy at nucleotide resolution and serving as a proxy for translation in the evaluated sample. Ribosome profiling has successfully underlined the identification and workings of differential splicing (Reixachs-Solé et al. 2020), microproteins (Weaver et al. 2019), and drug mechanisms (Chu and Pelletier 2018; Wolfe et al. 2014).

Similar to the output generated by other next-generation sequencing techniques, ribosome profiling results in complex data. Traditionally, processing of this data involves various manual processing steps that aggregates the data into curated features. However, ribosome profiling data has data set dependant biases and variance that are caused by both biological factors (e.g., tissue type, cell lines vs. tissue samples) and technical factors (e.g., translation inhibitors, lab protocols) (Mudge et al. 2022). As existing tools do not capture data set specific correlations, tools that delineate translated open reading frames (ORFs) are bound to suffer from low performances.

In this manuscript we introduce a custom transformer network, dubbed RIBO-former, for detecting translated ORFs using ribosome profiling next-generation sequencing data. RIBO-former features three distinct improvements over existing approaches. First, deep learning methodologies feature automated feature extraction, allowing us to apply minimal feature pre-processing, omitting steps that potentially constitute a loss of information or introduce biases. Second, our tool fine-tunes on independent data sets, allowing it to capture dataset-dependent correlations. Third, we apply attention-based neural networks, which are well suited to parse read information along variable-length RNA transcripts. These have been reported to achieve state-of-the-art performances on a variety of learning problems. We outline and evaluate multiple approaches to present ribosome profiling information into the transformer model. Through benchmarking, we shows RIBO-former to outperform existing tools.

Material and Methods

RIBO-former is created to map the translatome at single-nucleotide resolution of samples using ribosome profiling data . To simplify the experimental set-up, we train a model to detect active translation initiation sites (TISs), from which the open reading frame can be derived. RIBO-former processes ribosome profiling data along a full transcript (Figure 1). No pre-processing of other types of data, such as start codon or ORF information, are used to curate features or build a candidate ORF library. Instead, all positions on the transcriptome are evaluated using only information from mapped ribosome protected fragments. Mapping of the ribosome profiling data is performed using STAR (Dobin and Gingeras 2015) and cutadapt (Martin 2011).

Input data generation

Read lengths ranging from 20 to 40 nucleotides are included in the data. To allow computation with a transformer-based

*To whom correspondence should be addressed. email: clauwaer@umich.edu

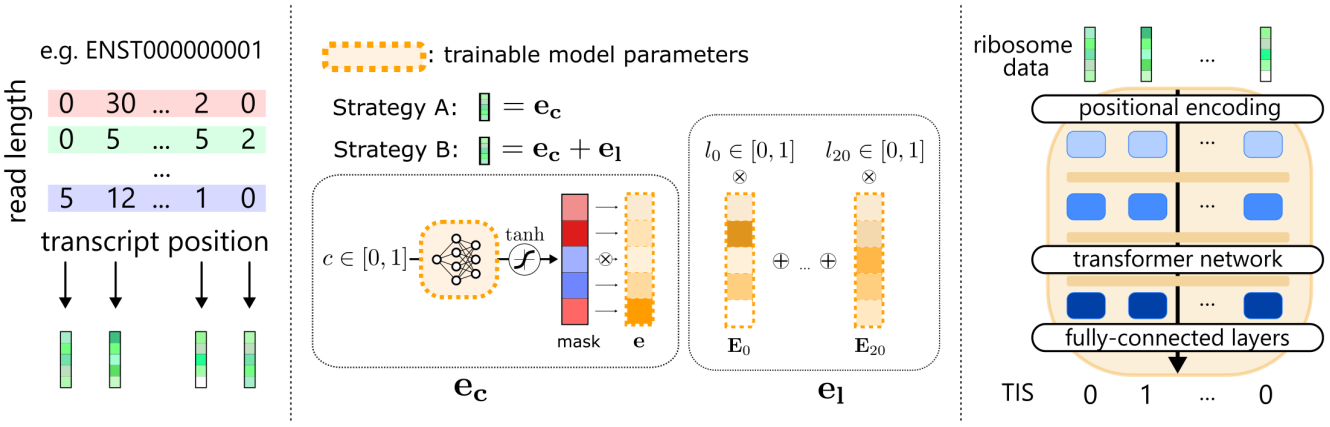

Figure 1: **Overview of RIBO-former pipeline.**(left) RIBO-former processes ribosome reads along a transcript region and generates vector embeddings for each transcript position based on the number of mapped reads and read length fractions. (middle) Two approaches are evaluated for the construction of input embeddings. One approach creates an input vector from the read counts at each position (A), while a second approach seeks to include read length information (B) (right) Illustration of the RIBO-former model. Expressed translation initiation sites (TISs) are predicted as a proxy for coding sequences.

architecture, vector representations are used that capture the occupancy of mapped reads by read length for each position (Figure 1). For a given position, the vector embedding $\mathbf{e_c}$ is obtained from the read count c using a set of feed-forward layers ϕ. Reads are only mapped to a single position on the transcript (e.g., by their 5'-end). Existing tools use a similar methodology, but apply various tools (e.g., Plastid (Dunn and Weissman 2016) and RiboWaltz (Lauria et al. 2018)) to offset mapped reads as a function of their read length. Read counts are normalized across the transcript for numerical stability.

$$\mathbf{e_c} = \mathbf{e} \odot \tanh(\phi(c)), \tag{1}$$

with $c \in [0,1]$, $\phi : \mathbb{R}^1 \rightarrow \mathbb{R}^h$, and $\mathbf{e} \in \mathbb{R}^h$. h is a hyperparameter of the model indicating the input dimension.

In this paper, we explore the inclusion of ribosome read length information as part of the information applied to determine TIS locations. For a given transcript position, $\mathbf{e_l}$ is calculated using the read length fractions $\mathbf{l}$ following the equation:

$$\mathbf{e_l} = \sum_{i=0}^{21} \mathbf{E}_i * l_i, \tag{2}$$

with $\mathbf{E} \in \mathbb{R}^{21 \times h}$ and $\mathbf{l} \in [0,1]^{21}$, where $\sum \mathbf{l} = 1$. The matrix $\mathbf{E}$ incorporates vector embeddings for read lengths 20–40 and is optimized as part of the training process. Note that ribosome data by read length is sparse and the majority of values in $\mathbf{l}$ are 0.

Table 1: **Data sets and their corresponding publications used in this study.** Only a single data set (SRR entry) is used from each study. CHX: Cycloheximide, LTM: Lactimidomycin, HRR: Harringtonine, CHL: Chloramphenicol

Author	SRR entry	Treatment	Mapped Reads
Benchmarking			
Ji et al.	SRR1802129	CHX	2.84E+06
Calviello et al.	SRR2433794	CHX	3.56E+07
Gawron et al.	SRR2732970	LTM	1.95E+08
Gawron et al.	SRR2733100	CHX	2.46E+08
Raj et al.	SRR2954800	HRR	3.87E+06
Martinez et al.	SRR8449577	CHX	1.19E+07
Chen et al.	SRR9113067	–	6.04E+06
Gaertner et al.	SRR11005875	CHX	9.75E+06
Pre-training			
Stern-Ginossar et al.	SRR592960	LTM	4.69E+06
Gonzalez et al.	SRR1562539	CHX	1.43E+07
Rubio et al.	SRR1573939	CHX	8.12E+07
Werner et al.	SRR1610244	CHX	1.43E+07
Tanenbaum et al.	SRR1976443	CHX	4.40E+07
Zur, Aviner, and Tuller	SRR2536856	CHX	1.74E+07
Loayza-Puch et al.	SRR2873532	CHL	1.93E+07
Bencun et al.	SRR3575904	HRR	8.47E+06

Model architecture and optimization

Continuing upon our previous work on detecting TISs using transcript sequence information (Clauwaert et al. 2023), we utilize an identical model framework for RIBO-former. Hyperparameter selection was performed using the data set featuring the most mapped reads (SRR2733100), where the optimal model features 212K weights (Algorithm 1). An important building block is a recent innovation, dubbed the performer, in calculating full attention introduced by Choromanski et al. (Choromanski et al. 2021), allowing long-range attention spanning full transcripts featuring tens of thousands of positions. The binary cross-entropy loss is used to optimize the model.

Table 2: **RIBO-former performances for different input token and optimization strategies**. Scores are calculated on the test set after selection of the model with the minimum validation loss. The area under the receiver operating characteristic curve (ROC) and area under the precision-recall curve (PR) are given. Strategy A has no read length information and generates input tokens by counting the total of reads at each position by taking the 5' position, or offsetting reads Plastid and RiboWaltz. Strategy B is calculated incorporating read length information. As strategy B performs optimal, this approach has been selected to evaluate pre-training of the model using a (self-)supervised learning objective.

	No pre-training								**Self-supervised**		**Supervised**	
	5' (A)		**Plastid (A)**		**RiboWaltz (A)**		**5' (B)**		**5' (B)**		**5' (B)**	
	ROC	**PR**	**ROC**	**PR**	**ROC**	**PR**	**ROC**	**PR**	**ROC**	**PR**	**ROC**	**PR**
SRR1802129	0.938	0.021	0.935	0.014	0.941	0.021	0.945	0.022	0.946	0.021	0.948	0.024
SRR2433794	0.965	0.094	0.964	0.083	0.966	0.091	0.968	0.104	0.966	0.104	0.969	0.112
SRR2732970	0.963	0.161	0.965	0.156	0.964	0.154	0.969	0.211	0.970	0.221	0.972	0.233
SRR2733100	0.965	0.161	0.964	0.145	0.964	0.151	0.970	0.217	0.969	0.220	0.972	0.235
SRR2954800	0.935	0.039	0.931	0.031	0.932	0.032	0.937	0.040	0.935	0.035	0.939	0.045
SRR8449577	0.955	0.072	0.954	0.058	0.955	0.064	0.956	0.075	0.959	0.086	0.961	0.094
SRR9113067	0.943	0.018	0.942	0.016	0.943	0.017	0.943	0.019	0.945	0.017	0.950	0.026
SRR11005875	0.967	0.073	0.966	0.072	0.967	0.072	0.967	0.080	0.968	0.078	0.971	0.092

Algorithm 1 RIBO-former network architecture. Given are the different layers, hyperparameter titles, optimal selection with bias term (italics) and total weights (bold).

RIBO-former | **211,964**
- **Read Count** | **21,546**
  - **Linear** | $1 \times \text{dim}$ | $1 \times 42 + \textit{42}$ | **84**
  - **Linear** | $\text{dim} \times \text{dim} * 6$ | $42 \times 252 + \textit{252}$ | **10,836**
  - **Linear** | $\text{dim} * 6 \times \text{dim}$ | $252 \times 42 + \textit{42}$ | **10,626**
- **Read Count Embedding** | $1 \times \text{dim}$ | 1×42 | **42**
- **Read Length Embedding** | read lengths $\times$ dim | 21×42 | **882**
- **Pos. Embedding** | fixed pos. embeddings | **0**
- **Performer** | **185,712**
  - **Layer** ($\times$ depth | 6) | **30,952**
    - **Layer norm** | $\text{dim} \times 2$ | $42 \times 2 + \textit{2}$ | **86**
    - **Attention head** ($\times$ n_head | 6) | **2,064**
      - $\mathbf{W_Q}$ | $\text{dim} \times \text{dim\_head}$ | $42 \times 16 + \textit{16}$ | **688**
      - $\mathbf{W_K}$ | $\text{dim} \times \text{dim\_head}$ | $42 \times 16 + \textit{16}$ | **688**
      - $\mathbf{W_V}$ | $\text{dim} \times \text{dim\_head}$ | $42 \times 16 + \textit{16}$ | **688**
      - $\mathbf{W_o}$ | $\text{dim\_head} * \text{n\_head} \times \text{dim}$ | $96 \times 42 + \textit{42}$ | **4,074**
    - **Layer norm** | $\text{dim} \times 2$ | $42 \times 2 + \textit{2}$ | **86**
    - **Linear** | $\text{dim} \times \text{dim} * 4$ | $42 \times 168 + \textit{168}$ | **7,224**
    - **Linear** | $\text{dim} * 4 \times \text{dim}$ | $168 \times 42 + \textit{42}$ | **7,098**
- **Linear** | $\text{dim} \times \text{dim} * 2$ | $42 \times 84 + \textit{84}$ | **3,612**
- **Linear** | $\text{dim} \times 2$ | $84 \times 2 + \textit{2}$ | **170**

Data selection and evaluation

A myriad of human ribosome profiling data sets were used covering a variety of tissues and treatment methods.The full transcriptome constitutes of 251,121 transcripts, with a total of 431,011,438 transcript positions. the annotated TISs from Ensembl GRCh38 version 107 are utilized as the positive set. In contrast to the annotations used to label the data, not all annotated TIS are expressed in each of the data sets. Therefore, a substantial set of positively annotated TISs cannot be predicted as no expression signal is present. While this affects the maximum performance tools can achieve on specific data sets, it does not affect our ability to compare tools based on their respective performances. Indeed, higher performances still indicate one tool to have a better capacity at detecting TISs from the ribosome profiling data. The training, validation and test sets are constructed from transcripts grouped per chromosome to ensure all transcript isoforms are present within the same set. Models are trained on the training set until a minimum loss on the validation set is reached, where results are acquired from the test set.

Results

Read length information can be leveraged to improve performances.

Today, the length of mapped reads serves as an important factor to determine the positioning of the ribosome protected fragments within the ribosome complex (Dunn and Weissman 2016). Existing methods evaluate the correlation between read length and the alignment of these reads with respect to the reading frame of existing coding sequences. This information is used by existing methods to offset the 5' mapping positions of existing reads when aggregating information from the variably sized mapped reads.

Here we show that the incorporating read length information, rather than aggregating this information by adjusting mapped positions, improves performances for RIBO-former (Table 2). Except for the few hundred weights used to compute the input vector, all approaches apply the same model architecture and, thus, number of model parameters ($\sim$220K). The difference between performances for different data sets reflects the effect of number of mapped reads (Table 1). Notably, as including read length information expands the number of features used to calculate the input vector embeddings, having a higher number of mapped reads influences the boost in performances when applying read length information (e.g., SRR2732970, SRRSRR2733100).

Pre-training improves both performances and training times

Ribosome profiling data is influenced by various factors, such as lab protocols, treatments, sample input material (cell types and/or species) and sequencing depth. By pre-training the model on a multitude of ribosome profiling data sets, it is able

Table 3: **Benchmark performances on detecting annotated coding sequences.**. Previous tools only evaluate a select number of ORFs/TISs, where a one-by-one comparison between these tools and RIBO-former is given. ROC: area under the receiver operating characteristic curve, PR: area under the precision-recall curve.

			PRICE		RIBO-former				RiboTish		RIBO-former	
Data set	**ORFs**	**CDSs**	**ROC**	**PR**	**ROC**	**PR**	**ORFs**	**CDSs**	**ROC**	**PR**	**ROC**	**PR**
SRR1802129	6,417	352	0.786	0.159	0.968	0.639	646,112	50,144	0.583	0.105	0.932	0.527
SRR2433794	42,978	2,534	0.693	0.185	0.961	0.639	1,020,518	69,479	0.556	0.078	0.941	0.558
SRR2732970	109,391	7,478	0.635	0.276	0.965	0.720	1,039,658	69,286	0.572	0.078	0.915	0.526
SRR2733100	122,202	4,735	0.562	0.153	0.977	0.694	1,002,275	67,270	0.565	0.077	0.919	0.535
SRR2954800	7,654	637	0.726	0.170	0.935	0.604	442,725	34,448	0.532	0.085	0.931	0.544
SRR8449577	15,520	1,027	0.728	0.207	0.958	0.650	842,566	60,171	0.574	0.089	0.945	0.568
SRR9113067	11,944	457	0.781	0.195	0.967	0.565	759,182	55,710	0.553	0.086	0.924	0.472
SRR11005875	13,175	1,477	0.721	0.246	0.943	0.678	1,006,709	71,656	0.566	0.085	0.944	0.563

			Rp-Bp		RIBO-former				Ribotricer		RIBO-former	
Data set	**ORFs**	**CDSs**	**ROC**	**PR**	**ROC**	**PR**	**ORFs**	**CDSs**	**ROC**	**PR**	**ROC**	**PR**
SRR1802129	269,574	28,401	0.579	0.143	0.928	0.578	238,095	23,206	0.569	0.111	0.966	0.709
SRR2433794	453,936	40,037	0.564	0.108	0.945	0.621	506,515	33,181	0.594	0.078	0.974	0.692
SRR2732970	485,714	40,407	0.622	0.118	0.921	0.583	487,689	28,823	0.557	0.065	0.971	0.687
SRR2733100	472,020	39,761	0.611	0.115	0.924	0.593	171,597	4,022	0.603	0.034	0.964	0.447
SRR2954800	237,475	23,265	0.536	0.112	0.922	0.556	155,014	15,248	0.586	0.122	0.959	0.679
SRR8449577	369,043	34,404	0.578	0.125	0.946	0.622	363,661	28,043	0.592	0.093	0.973	0.706
SRR9113067	372,811	34,991	0.547	0.109	0.917	0.507	195,331	15,107	0.593	0.099	0.962	0.646
SRR11005875	446,978	41,421	0.566	0.113	0.947	0.627	439,689	34,384	0.582	0.091	0.971	0.694

to discover general RIBO-seq correlations that are shared between different data sets. A selection of eight complementary ribosome profiling experiments were selected to pre-train the model (Table 1).

Two types of pre-trained models have been evaluated. The first set of pre-trained models is optimized to detect TISs (supervised), following an identical learning objective as outlined above. The second set of models is optimized using a self-supervised learning setting, where labels are derived from the input data. This is a popular approach in natural language processing, producing models capable of handling a variety of tasks, In this setting, random locations of the input are masked, where the model is tasked to impute the presence of ribosome reads at these positions. Performances are evaluated for models incorporating read length information.

The results show that pre-training a model on a variety of data improves performances for the supervised learning objective (Table 2) Notably, training times are drastically reduced on all evaluated data sets, where fine-tuning of models on the pre-trained model reaches convergence on the validation set after only one training epoch (versus 10 epochs).

RIBO-former outperforms existing tools by a large margin

Several tools exist that utilize ribosome profiling data in various manners to delineate translated ORFs. Tools like RiboTaper (Calviello et al. 2016), riboHMM (Raj et al. 2016b), and Scikit-ribo (Fang et al. 2018) are packages that combine both RNA-seq and Ribo-seq data to achieve this goal. Other software, such as PROTEOFORMER (Crappé et al. 2015) and RiboTISH (Zhang et al. 2017), can leverage information from the combination of cycloheximide and lactimidomycin

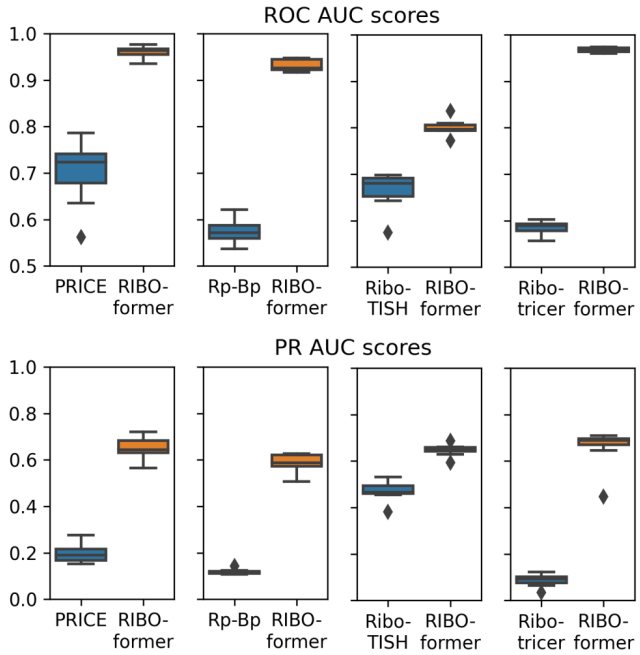

Figure 2: **Benchmark performances on detecting annotated coding sequences.** Figures are derived from the data listed in Table 3. Previous tools only evaluate a select number of ORFs, where a one-by-one comparison between these tools and RIBO-former is given. ROC: area under the receiver operating characteristic curve, PR: area under the precision-recall curve.

or harringtonine treated data sets.

In this study, we benchmark RIBO-former with PRICE (Erhard et al. 2018), Rp-Bp (Malone et al. 2017), RiboTISH (Zhang et al. 2017), and Ribotricer (Choudhary, Li, and D. Smith 2020), using the annotated coding sequences by Ensembl as the positive set. These tools were selected based on various factors: previously reported benchmarks, the presence of continued support for a tool through software updates and code maintenance, and their adoption by the community (Xiao et al. 2018; Choudhary, Li, and D. Smith 2020; Malone et al. 2017). RIBO-former predicts translated TISs along the full transcriptome, from which translated ORFs are derived. Unlike RIBO-former, existing methods are designed to only evaluate a subset of candidate ORFs. As such, only one-by-one comparisons between each tool and RIBO-former are possible, where only the positions included by each tool are used to calculate the performance scores. Performances between existing tools cannot be compared, as different sets of evaluated ORFs decide the complexity of the benchmark and, therefore, the resulting performance metric. Importantly, post-processing steps that further filter the output predictions of all tools have been omitted, as these steps further limit the number of evaluated samples and do not represent the predictive performance of the statistical methods incorporated for each tool. Examples of such post-processing steps include the selection of only the longest ORFs on each transcript.

We show RIBO-former to outperform all other tools on all evaluated data sets (Table 3, Figure 2). Here, RIBO-former especially contrasts itself from existing approaches when it is evaluated on large sets featuring few positive samples.

Discussion

In this paper, we propose RIBO-former, an attention-based tool for detecting translated ORFs from ribosome profiling read information along the transcript. We show our method to substantially outperform previous tools, and believe our tool to bring various improvements. (i) The tool relies solely on the arrangement of ribosome-protected fragments to detect TISs. Other information, such as sequence information (e.g., start codon, stop codon) or information on the properties of the open reading frame (e.g., length, number of reads mapped) could perpetuate potential biases if used as input features of statistical methods. If necessary, filtering based on these properties can always be achieved as a post-processing step, rather than a pre-processing step. (ii) The tool omits pre-processing steps that re-map reads as a function of their read length. As not all read lengths map to a single reading frame, we know this step constitutes a loss of information or introduces bias. We show RIBO-former to achiev higher performances when parsing read length information as is. These results showcase the strength of deep learning approaches to do automated feature extraction. (iii) The mapping of the full transcriptome allows for modular post-processing steps where specific sites of interest are guaranteed to be evaluated. Filters based on number of reads per transcript, sequence information, and ORF properties can be applied on the full set in line with the objectives of the user. Furthermore, evaluating the full transcriptome facilitates future benchmarking. (iv) The application of state-of-the-art machine learning models and the incorporation of ribosome length information outperforms all previously designed tools for all evaluated data sets. Importantly, the tool was shown to work well with a variety of data, covering a wide range of sequencing depths, sample types, and antibiotic treatments applied. We find these observations to support the utility of the tool. Although a focus was set on the human genome, RIBO-former can be applied across different species, and future findings might even prove the inclusion of data sets from multiple species to be a valid strategy to create a pre-trained model.

When running RIBO-former on new data, it fine-tunes pre-trained models on two non-overlapping folds, where predictions are obtained on the test sets covering the full transcriptome. As such, the tool relies on the availability of a graphical processing unit from the user, which can serve as a limiting factor for the adoption of the tool. The use of pre-trained models alleviates the requirement of computational resources, as training times are reduced to a single epoch (15 min. on RTX3090). The accessibility of cloud computing solutions can further provide support for the application of graphical processing unit-powered algorithms in labs that are otherwise lacking support of local hardware.

Future work on RIBO-former will focus on further improving the tool through various readily-available post-processing steps. To illustrate, we found RIBO-former to suffer from from low accuracy for transcripts featuring a lower number of mapped reads, where the tool sometimes misses existing TISs by a few nucleotides. Post-processing steps can detect and correct likely inaccuracies by evaluating model scores and neighboring codon prevalence. Other post-processing strategies can introduce filters based on read count and ORF properties, such as start codon and ORF lengths.

RIBO-former has been designed to complement TIS transformer (Clauwaert et al. 2023), a tool previously designed by our team that applies transcript sequence information to delineate coding sequences on the transcriptome. With RIBO-former achieving high performances solely using ribosome profiling information, we believe the tool to be a promising new asset with which the precise detection of novel translated non-canonical ORFs can be achieved.

Data availability

All the data, scripts and model outputs are available for public use on GitHub (`https://github.com/jdcla/RIBO_former_paper`) and Zenodo (`https://doi.org/10.5281/zenodo.8059446`). The RIBO-former is also available on GitHub `https://github.com/jdcla/RIBO_former` and PyPI (`https://pypi.org/project/transcript-transformer/`).

Funding

The work presented was sponsored by Novo Nordisk Research Centre Oxford Ltd with the work being carried out jointly by Ghent University and Novo Nordisk employees. Novo Nordisk took part in the study design and overall supervision and guidance of the project, with a focus on the evaluation of biological relevance. Ghent University and University

of Michigan were responsible for the study design, technical development, testing and implementation of the deep learning model. J.R.P. acknowledges funding from the National Institutes of Health/National Cancer Institute (K08-CA263552-01A1), the Alexs Lemonade Stand Foundation Young Investigator Award (#21-23983), the St. Baldricks Foundation Scholar Award (#931638), the Musella Foundation for Brain Tumor Research, the DIPG/DMG Research Funding Alliance, and a Collaborative Pediatric Cancer Research Awards Program/Kids Join the Fight award (#22FN23).

Declaration of interests

The authors declare no competing interests.

References

Bencun, M.; Klinke, O.; Hotz-Wagenblatt, A.; Klaus, S.; Tsai, M.-H.; Poirey, R.; and Delecluse, H.-J. 2018. Translational profiling of B cells infected with the Epstein-Barr virus reveals 5′ leader ribosome recruitment through upstream open reading frames. *Nucleic Acids Research* 46(6):2802–2819.

Calviello, L.; Mukherjee, N.; Wyler, E.; Zauber, H.; Hirsekorn, A.; Selbach, M.; Landthaler, M.; Obermayer, B.; and Ohler, U. 2016. Detecting actively translated open reading frames in ribosome profiling data. *Nature Methods* 13(2):165–170.

Chen, J.; Brunner, A.-D.; Cogan, J. Z.; Nuñez, J. K.; Fields, A. P.; Adamson, B.; Itzhak, D. N.; Li, J. Y.; Mann, M.; Leonetti, M. D.; and Weissman, J. S. 2020. Pervasive functional translation of noncanonical human open reading frames. *Science* 367(6482):1140–1146.

Choromanski, K.; Likhosherstov, V.; Dohan, D.; Song, X.; Gane, A.; Sarlos, T.; Hawkins, P.; Davis, J.; Mohiuddin, A.; Kaiser, L.; Belanger, D.; Colwell, L.; and Weller, A. 2021. Rethinking Attention with Performers. *arXiv:2009.14794 [cs, stat]*.

Choudhary, S.; Li, W.; and D. Smith, A. 2020. Accurate detection of short and long active ORFs using Ribo-seq data. *Bioinformatics* 36(7):2053–2059.

Chu, J., and Pelletier, J. 2018. Therapeutic Opportunities in Eukaryotic Translation. *Cold Spring Harbor Perspectives in Biology* 10(6):a032995.

Clauwaert, J.; McVey, Z.; Gupta, R.; and Menschaert, G. 2023. TIS Transformer: Remapping the human proteome using deep learning. *NAR genomics and bioinformatics* 5(1):lqad021.

Crappé, J.; Ndah, E.; Koch, A.; Steyaert, S.; Gawron, D.; De Keulenaer, S.; De Meester, E.; De Meyer, T.; Van Criekinge, W.; Van Damme, P.; and Menschaert, G. 2015. PROTEOFORMER: Deep proteome coverage through ribosome profiling and MS integration. *Nucleic Acids Research* 43(5):e29.

Dobin, A., and Gingeras, T. R. 2015. Mapping RNA-seq Reads with STAR. *Current Protocols in Bioinformatics* 51(1):11.14.1–11.14.19.

Dunn, J. G., and Weissman, J. S. 2016. Plastid: nucleotide-resolution analysis of next-generation sequencing and genomics data. *BMC Genomics* 17(1):958.

Erhard, F.; Halenius, A.; Zimmermann, C.; L'Hernault, A.; Kowalewski, D. J.; Weekes, M. P.; Stevanovic, S.; Zimmer, R.; and Dölken, L. 2018. Improved Ribo-seq enables identification of cryptic translation events. *Nature Methods* 15(5):363–366.

Fang, H.; Huang, Y.-F.; Radhakrishnan, A.; Siepel, A.; Lyon, G. J.; and Schatz, M. C. 2018. Scikit-ribo Enables Accurate Estimation and Robust Modeling of Translation Dynamics at Codon Resolution. *Cell Systems* 6(2):180–191.e4.

Gaertner, B.; van Heesch, S.; Schneider-Lunitz, V.; Schulz, J. F.; Witte, F.; Blachut, S.; Nguyen, S.; Wong, R.; Matta, I.; Hübner, N.; and Sander, M. 2020. A human ESC-based screen identifies a role for the translated lncRNA LINC00261 in pancreatic endocrine differentiation. *eLife* 9:e58659.

Gawron, D.; Ndah, E.; Gevaert, K.; and Van Damme, P. 2016. Positional proteomics reveals differences in N-terminal proteoform stability. *Molecular Systems Biology* 12(2):858.

Gonzalez, C.; Sims, J. S.; Hornstein, N.; Mela, A.; Garcia, F.; Lei, L.; Gass, D. A.; Amendolara, B.; Bruce, J. N.; Canoll, P.; and Sims, P. A. 2014. Ribosome Profiling Reveals a Cell-Type-Specific Translational Landscape in Brain Tumors. *Journal of Neuroscience* 34(33):10924–10936.

Ingolia, N. T.; Ghaemmaghami, S.; Newman, J. R. S.; and Weissman, J. S. 2009. Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science* 324(5924):218–223.

Ji, Z.; Song, R.; Regev, A.; and Struhl, K. 2015. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *eLife* 4:e08890.

Lauria, F.; Tebaldi, T.; Bernabò, P.; Groen, E. J. N.; Gillingwater, T. H.; and Viero, G. 2018. riboWaltz: Optimization of ribosome P-site positioning in ribosome profiling data. *PLOS Computational Biology* 14(8):e1006169.

Loayza-Puch, F.; Rooijers, K.; Buil, L. C. M.; Zijlstra, J.; F. Oude Vrielink, J.; Lopes, R.; Ugalde, A. P.; van Breugel, P.; Hofland, I.; Wesseling, J.; van Tellingen, O.; Bex, A.; and Agami, R. 2016. Tumour-specific proline vulnerability uncovered by differential ribosome codon reading. *Nature* 530(7591):490–494.

Malone, B.; Atanassov, I.; Aeschimann, F.; Li, X.; Großhans, H.; and Dieterich, C. 2017. Bayesian prediction of RNA translation from ribosome profiling. *Nucleic Acids Research* 45(6):2960–2972.

Martin, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17(1):10–12.

Martinez, T. F.; Chu, Q.; Donaldson, C.; Tan, D.; Shokhirev, M. N.; and Saghatelian, A. 2020. Accurate annotation of human protein-coding small open reading frames. *Nature Chemical Biology* 16(4):458–468.

Mudge, J. M.; Ruiz-Orera, J.; Prensner, J. R.; Brunet, M. A.; Calvet, F.; Jungreis, I.; Gonzalez, J. M.; Magrane, M.; Martinez, T. F.; Schulz, J. F.; Yang, Y. T.; Albà, M. M.; Aspden,

J. L.; Baranov, P. V.; Bazzini, A. A.; Bruford, E.; Martin, M. J.; Calviello, L.; Carvunis, A.-R.; Chen, J.; Couso, J. P.; Deutsch, E. W.; Flicek, P.; Frankish, A.; Gerstein, M.; Hubner, N.; Ingolia, N. T.; Kellis, M.; Menschaert, G.; Moritz, R. L.; Ohler, U.; Roucou, X.; Saghatelian, A.; Weissman, J. S.; and van Heesch, S. 2022. Standardized annotation of translated open reading frames. *Nature Biotechnology* 40(7):994–999.

Raj, A.; Wang, S. H.; Shim, H.; Harpak, A.; Li, Y. I.; Engelmann, B.; Stephens, M.; Gilad, Y.; and Pritchard, J. K. 2016a. Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. *eLife* 5:e13328.

Raj, A.; Wang, S. H.; Shim, H.; Harpak, A.; Li, Y. I.; Engelmann, B.; Stephens, M.; Gilad, Y.; and Pritchard, J. K. 2016b. Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. *eLife* 5:e13328.

Reixachs-Solé, M.; Ruiz-Orera, J.; Albà, M. M.; and Eyras, E. 2020. Ribosome profiling at isoform level reveals evolutionary conserved impacts of differential splicing on the proteome. *Nature Communications* 11(1):1768.

Rubio, C. A.; Weisburd, B.; Holderfield, M.; Arias, C.; Fang, E.; DeRisi, J. L.; and Fanidi, A. 2014. Transcriptome-wide characterization of the eIF4A signature highlights plasticity in translation regulation. *Genome Biology* 15(10):476.

Stern-Ginossar, N.; Weisburd, B.; Michalski, A.; Le, V. T. K.; Hein, M. Y.; Huang, S.-X.; Ma, M.; Shen, B.; Qian, S.-B.; Hengel, H.; Mann, M.; Ingolia, N. T.; and Weissman, J. S. 2012. Decoding Human Cytomegalovirus. *Science* 338(6110):1088–1093.

Tanenbaum, M. E.; Stern-Ginossar, N.; Weissman, J. S.; and Vale, R. D. 2015. Regulation of mRNA translation during mitosis. *eLife* 4:e07957.

Weaver, J.; Mohammad, F.; Buskirk, A. R.; and Storz, G. 2019. Identifying Small Proteins by Ribosome Profiling with Stalled Initiation Complexes. *mBio* 10(2):e02819–18.

Werner, A.; Iwasaki, S.; McGourty, C. A.; Medina-Ruiz, S.; Teerikorpi, N.; Fedrigo, I.; Ingolia, N. T.; and Rape, M. 2015. Cell-fate determination by ubiquitin-dependent regulation of translation. *Nature* 525(7570):523–527.

Wolfe, A. L.; Singh, K.; Zhong, Y.; Drewe, P.; Rajasekhar, V. K.; Sanghvi, V. R.; Mavrakis, K. J.; Jiang, M.; Roderick, J. E.; Van der Meulen, J.; Schatz, J. H.; Rodrigo, C. M.; Zhao, C.; Rondou, P.; de Stanchina, E.; Teruya-Feldstein, J.; Kelliher, M. A.; Speleman, F.; Porco, J. A.; Pelletier, J.; Rätsch, G.; and Wendel, H.-G. 2014. RNA G-quadruplexes cause eIF4A-dependent oncogene translation in cancer. *Nature* 513(7516):65–70.

Xiao, Z.; Huang, R.; Xing, X.; Chen, Y.; Deng, H.; and Yang, X. 2018. De novo annotation and characterization of the translatome with ribosome profiling data. *Nucleic Acids Research* 46(10):e61.

Zhang, P.; He, D.; Xu, Y.; Hou, J.; Pan, B.-F.; Wang, Y.; Liu, T.; Davis, C. M.; Ehli, E. A.; Tan, L.; Zhou, F.; Hu, J.; Yu, Y.; Chen, X.; Nguyen, T. M.; Rosen, J. M.; Hawke, D. H.; Ji, Z.; and Chen, Y. 2017. Genome-wide identification and differential analysis of translational initiation. *Nature Communications* 8(1):1749.

Zur, H.; Aviner, R.; and Tuller, T. 2016. Complementary Post Transcriptional Regulatory Information is Detected by PUNCH-P and Ribosome Profiling. *Scientific Reports* 6(1):21635.