DATA NOTE

REVISED Ribosome profiling of HEK293T cells overexpressing codon optimized coagulation factor IX [version 2; peer review: 3 approved]

Aikaterini Alexaki[1*], Jacob Kames[1*], Gaya K. Hettiarachchi[1], John C. Athey[1], Upendra K. Katneni[1], Ryan C. Hunt[1], Nobuko Hamasaki-Katagiri[1], David D. Holcomb[1], Michael DiCuccio[2], Haim Bar[3], Anton A. Komar[4], Chava Kimchi-Sarfaty[1]

[1]Center for Biologics Evaluation and Research, Food and Drug Administration, USA, Silver Spring, MD, 20993, USA
[2]National Center of Biotechnology Information, National Institutes of Health, USA, Bethesda, MD, 20892, USA
[3]Department of Statistics, University of Connecticut, Storrs, CT, 06269, USA
[4]Center for Gene Regulation in Health and Disease, Cleveland State University, Cleveland, OH, 44115, USA

[*] Equal contributors

v2 **First published:** 10 Mar 2020, **9**:174
https://doi.org/10.12688/f1000research.22400.1

Latest published: 21 Sep 2020, **9**:174
https://doi.org/10.12688/f1000research.22400.2

Abstract

Ribosome profiling provides the opportunity to evaluate translation kinetics at codon level resolution. Here, we describe ribosome profiling data, generated from two HEK293T cell lines. The ribosome profiling data are composed of Ribo-seq (mRNA sequencing data from ribosome protected fragments) and RNA-seq data (total RNA sequencing). The two HEK293T cell lines each express a version of the *F9* gene, both of which are translated into identical proteins in terms of their amino acid sequences. However, these *F9* genes vary drastically in their codon usage and predicted mRNA structure. We also provide the pipeline that we used to analyze the data. Further analyzing this dataset holds great potential as it can be used i) to unveil insights into the composition and regulation of the transcriptome, ii) for comparison with other ribosome profiling datasets, iii) to measure the rate of protein synthesis across the proteome and identify differences in elongation rates, iv) to discover previously unidentified translation of peptides, v) to explore the effects of codon usage or codon context in translational kinetics and vi) to investigate cotranslational folding. Importantly, a unique feature of this dataset, compared to other available ribosome profiling data, is the presence of the *F9* gene in two very distinct coding sequences.

Keywords

Ribosome profiling, codon optimization, Ribo-seq, RNA-seq, translation kinetics, codon usage, codon pair usage, protein therapeutics

Open Peer Review

Reviewer Status ✓ ✓ ✓

Invited Reviewers

	1	2	3
version 2 (revision) 21 Sep 2020	✓	✓ report	✓ report
version 1 10 Mar 2020	? report	? report	? report

1. **Stefano Biffo**, INGM, National Institute of Molecular Genetics, “Fondazione Romeo ed Enrica Invernizzi”, Milan, Italy
   Riccardo Rossi, INGM, Milan, Italy
2. **Jordan Berg**, University of Utah, Salt Lake City, USA
3. **James Collawn**, University of Alabama at Birmingham, Birmingham, USA

Rafal Bartoszewski, Medical University of Gdansk, Gdansk, Poland

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Chava Kimchi-Sarfaty (Chava.kimchi-sarfaty@fda.hhs.gov)

Author roles: Alexaki A: Data Curation, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Kames J**: Data Curation, Software; **Hettiarachchi GK**: Conceptualization, Formal Analysis, Investigation; **Athey JC**: Formal Analysis, Software, Visualization; **Katneni UK**: Formal Analysis, Investigation; **Hunt RC**: Conceptualization, Investigation; **Hamasaki-Katagiri N**: Data Curation, Formal Analysis; **Holcomb DD**: Data Curation, Formal Analysis; **DiCuccio M**: Formal Analysis; **Bar H**: Data Curation, Formal Analysis; **Komar AA**: Conceptualization; **Kimchi-Sarfaty C**: Conceptualization, Formal Analysis, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by funds from the U.S. Food and Drug Administration Chief Scientist grant and in part by an appointment to the Research Participation Program at the Center for Biologics Evaluation and Research administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the U.S. Food and Drug Administration (C.K.-S.). Additionally, this work was in part supported by the National Institutes of Health grant HL151392 (A.A.K.).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Alexaki A, Kames J, Hettiarachchi GK *et al.* **Ribosome profiling of HEK293T cells overexpressing codon optimized coagulation factor IX [version 2; peer review: 3 approved]** F1000Research 2020, **9**:174 https://doi.org/10.12688/f1000research.22400.2

First published: 10 Mar 2020, **9**:174 https://doi.org/10.12688/f1000research.22400.1

REVISED **Amendments from Version 1**

The ribosome profiling data analysis pipeline described in the manuscript has been extensively updated in terms of its usability and composition of tools. We have updated the alignment tool from Tophat to HISAT2, changed the code from Python 2.7 to Python 3.7 and updated many of the scripts to make the code more readable and generate error messages to assist in debugging. Furthermore, we have bundled all of the scripts involved in the pipeline into two more easily run bash scripts to aid in usability. The GitHub documentation and usage notes have also been updated to more explicitly describe versions of assemblies and tools used in the pipeline. We have updated the figures in the paper with data generated from the new pipeline and have added an additional statistic in the data validation section, the Spearman's rank correlation coefficient.

Any further responses from the reviewers can be found at the end of the article

Introduction

The ribosome profiling (footprinting) technique has only been around for a decade[1] but has already contributed tremendously to our understanding of translation efficiency and kinetics. Initially developed to systematically monitor protein translation in yeast[1], it has since been adapted to work in a range of organisms[2,3] and to tackle a variety of questions. Ribosome profiling data typically consist of a set of sequences of ribosome protected fragments (RPF), designated as Ribo-seq data, which is accompanied by sequences from total RNA (RNA-seq). The availability of Ribo-seq and RNA-seq data from the same sample provides a treasure trove of information, enabling quantitative study of translation efficiency, rate and kinetics of every mRNA sequence in the pool[4]. Given that these sequences cover the entire transcriptome, and also include tRNA and rRNA, typically only a fraction of the data is presented and constructively used, within its initial publication. Further analyses, and comparisons of different ribosome profiling datasets can yield significant new information.

We recently conducted a ribosome profiling study to examine the translation kinetics of blood coagulation factor IX[5], a protein with great pharmaceutical interest. Two human embryonic kidney 293T (HEK293T) cell lines were lentivirally transduced, one with the wild type (WT) version of the gene and one with a codon optimized (CO) *F9*[5]. Codon optimization is a widely used technique that aims at increasing the protein expression levels by replacing multiple codons within a coding sequence with synonymous ones. In doing so, the amino acid sequence of the protein remains unaltered, therefore these changes were assumed to be inconsequential for the structure and function of the protein. However, this is not always true; through our ribosome profiling study, we described that these synonymous changes drastically altered translational kinetics and led to protein conformational changes[5].

The translational kinetics of the *F9* variants, along with the control genes, *GAPDH* and *ACTB*, were analyzed in detail in the original publication[5]. Similarly, any other gene of interest can be investigated in this dataset in terms of their rate of synthesis and translational kinetics; genes in the entire transcriptome can be compared to each other. Since there are several other HEK293T ribosomal profiling datasets available, these could be used to examine the reproducibility of the results[6]. Furthermore, by looking into ribosome profiling datasets from other cell types, such as other human cells[7] and/or across species, it would be valuable to examine whether a given gene maintains the same translation kinetics or if there are significant differences that could reflect on the conformation of the protein. Clearly, since a rather large inter-experiment variation is expected, the accumulation of several ribosome profiling databases would be very useful for this type of analysis.

Innovative computational approaches of analyzing ribosome profiling data have led to the identification of novel CDSs that lead to the production of previously unidentified peptides and variants of known proteins[8]. Such coding sequences may be found in what is typically designated as untranslated regions (UTRs) of the mRNA, particularly the 5'UTRs, and may originate from non-AUG start sites[9–11]. However, such approaches have not been applied yet to this dataset and it would be intriguing to see if they could lead to new discoveries[12]. Importantly, since the genome of the HEK293T used to generate this dataset contains part of lentiviral vector and the cytomegalovirus (CMV) promoter to drive expression of *F9*, it would be interesting to examine whether any part of this sequence is actively translated. These analyses may be particularly insightful in studies of immunogenicity.

Further analysis of this dataset will help elucidate the effect of codon usage, codon context and possible other factors in translational kinetics. By looking at the global rate in which each codon is translated, and examining adjacent sequences on a transcriptome level, it may be possible to predict translational kinetics of recombinant genes and to make inferences on whether cotranslational folding may be affected. This may be particularly important in gene therapy applications where the cell type expressing the gene of interest may be different from the naturally expressing cells, e.g. expression of coagulation factor VIII from hepatocytes in gene therapy. A recent study in yeast[13] showed promising results in this direction; however, increasing availability of ribosome profiling datasets from other cell types will allow further comparisons. A unique feature of this dataset that may be pivotal in these types of studies is the presence of *F9* in two genes with very different codon usage.

Materials and methods

Plasmid/vector construction

WT (RefSeq NM_000133.3) and CO (accessible at https://github.com/FDA/Ribosome-Profiling F9_opt1_construct_100bpUTRs.fasta)[14] *F9* ORFs were sub-cloned into pcDNA3.1/V5-His-TOPO

(Invitrogen/Life Technologies) according to manufacturer's instructions to generate pcDNA3.1-*F9*-V5-His plasmids. Each fusion construct (WT*F9*-V5-His and CO*F9*-V5-His) was subcloned into a lentiviral vector pTK642 (gift from Dr. Kafri, University of North Carolina at Chapel Hill) at the Pacl/Sfil site.

Cell cultures and lentiviral transduction

Human embryonic kidney cells (HEK293T; ATCC) were grown in Dulbecco's Modified Eagle Medium (Quality Biological, Inc) with 1% L-glutamine (Quality Biological), 1% penicillin-streptomycin (Hyclone) and 10% fetal bovine serum (Quality Biological) at 37°C in 5% CO_2. HEK293T cells stably expressing WT or CO FIX were established following transduction with lentiviral vectors, as previously described[15].

An equivalent number of cells were plated in T-flasks and supplemented with 10 ng/ml of Vitamin K3, one day prior to all experiments. The culture medium was replaced with Opti-MEM Reduced Serum Medium (Life Technologies) at approximately 80–90% cell confluency and cells were harvested after an additional 24 hours of incubation.

Ribosome profiling

Ribosome profiling was conducted as described previously[7] using the Illumina TruSeq Ribo Profile (Mammalian) Kit according to manufacturer's instructions with modifications in harvest, RNA isolation/purification (isopropanol isolation used to improve the yield) and ribosome protected fragments size selection (~20–32 nt). During harvest, media was carefully removed, and cells were immediately flash-frozen. All equipment used from hence forth was pre-chilled. Cells were quickly scraped into 1 ml of ice-cold lysis buffer (5X Mammalian Polysome Buffer, 10% Triton-X100, 100 mM DTT, DNase I, Nuclease-free water) and homogenized on ice by passing through a 26G needle 10 times. Lysate was then spun at 4°C for 10 minutes at 20,000 × g. Supernatant was aliquoted into cryovials and immediately frozen in liquid nitrogen for future use. Samples were sequenced using Illumina HiSeq 2500.

The complete ribosome profiling pipeline analysis is described in Figure 1: Sequencing data were pre-processed and aligned as described by Alexaki *et al.*[5] as well as the step by step guide found in the README.txt accessible on GitHub.

RPF sequences were analyzed based on fragment length (Figure 2a), alignment distribution between coding sequences (CDSs) and 5'- and 3'-UTRs (Figure 2b), triplet periodicity (Figure 3a) and reading frame (Figure 3b). RPF fragments 20–22 nt and 27–29 nt in length were used for further analysis with a P-site offset of 12 nucleotides from the 5' end of the fragment. Pearson and Spearman correlations were used to evaluate the reproducibility between replicates using a common subset of moderately to highly expressed genes (reads per kilobase of transcript per million mapped reads, $RPKM_{CDS}$ ≥10) and considering reads with the ribosome A site annotated at least 20 nt downstream of the coding sequence start codon (Table 1). Both Pearson and Spearman coefficients show strong correlation between experimental replicates.

Dataset validation

The quality of the sequencing files is presented in Table 2. A pipeline was created to process the data (Figure 1). A number of steps allow for validation of the data and confirmation of their quality. The fragment length distributions for the whole genome were plotted, indicating that the vast majority of the fragments from the Ribo-seq data are either 20–21 or 27–28 nucleotides in length (Figure 2a), and as expected the RNA-seq data have a more flat distribution. The distribution of the Ribo-seq data in the UTRs and CDSs of the mRNA was also plotted. As expected, most of the sequences aligned within the CDSs (Figure 2b), while a smaller fraction of the RNA-seq data aligned with the CDSs. It should be noted that as the 3' UTR, and 5' UTR are typically shorter in length than the CDSs, it is not surprising that about 60% of the RNA-seq data align with the CDSs (Figure 2b). In addition, Ribo-seq data exhibit periodicity, characteristic of the RPFs (Figure 3a and Figure 3b), which is not observed in the RNA-seq data (Figure 3b). In accordance with previously published data[16], we can infer that the 5′-most peaks in (Figure 3a) represent ribosomes with the start codon in the P site and the second codon in the A site, for both large and short fragments. Very tight correlation between the experiments, both for Ribo-seq and RNA-seq data, supports the reproducibility of the results (Table 1).

Data records

Sequencing for 3 replicates of RNA-seq and Ribo-seq of HEK293T cells stably expressing WT and CO FIX was performed by Eurofins Genomics (Louisville, KY, USA), resulting in 12 raw data files (3 WT and 3 CO *F9* for both Ribo-seq and RNA-seq) in FASTQ format. Raw data are accessible at the NCBI Sequence Read Archive (SRA) under BioProject accession PRJNA591214. File names, SRA accession numbers (experiment and sample) and descriptions of data are summarized below in Table 3.

Usage notes

The custom ribosome profiling analysis pipeline has been deposited in GitHub in the FDA/Ribosome-Profiling directory[14]. Raw data files may be accessed from SRA and downloaded to the './Ribosome_profiling/Raw_data/X/' folder. In our descriptions and instructions, 'X' is replaced with 'S12', but the user may choose any designation they prefer. Detailed instructions for running the data analysis pipeline are included in the 'README.txt' file.

Execution of the pipeline requires the following tools (version tested) be installed on the user's system: Python (3.7.6) (https://www.python.org) (Python Software Foundation, Wilmington, DE, USA) and modules pysam (0.15.3) (https://github.com/pysam-developers/pysam) and biopython (1.77) (https://biopython.org/), GFF Utilities (gffread v0.12.1) (http://ccb.jhu.edu/software/stringtie/gff.shtml) (Johns Hopkins University, Baltimore, MD,

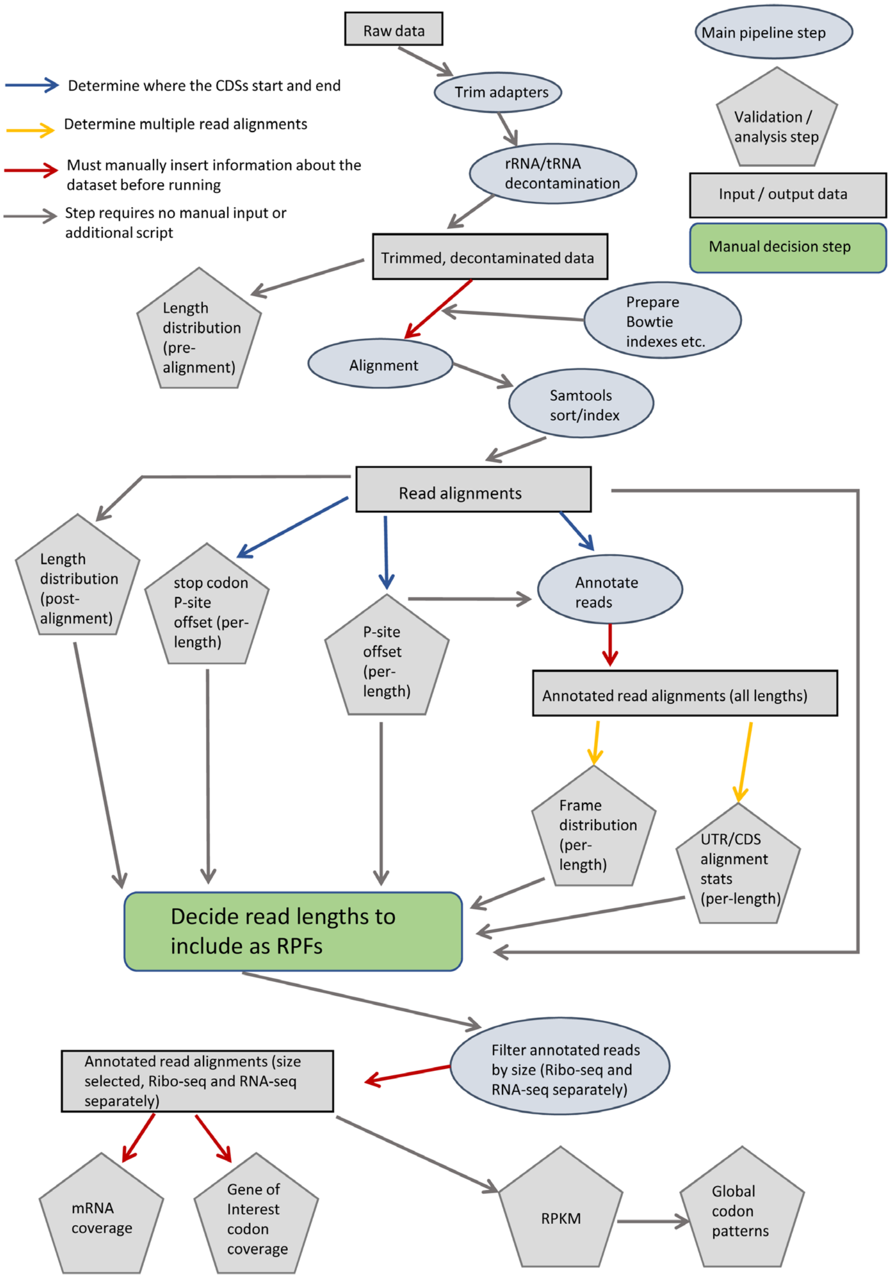
Raw data
Main pipeline step
Determine where the CDSs start and end
Determine multiple read alignments
Must manually insert information about the dataset before running
Step requires no manual input or additional script
Trim adapters
Validation / analysis step
rRNA/tRNA decontamination
Input / output data
Manual decision step
Trimmed, decontaminated data
Length distribution (pre-alignment)
Prepare Bowtie indexes etc.
Alignment
Samtools sort/index
Read alignments
Length distribution (post-alignment)
stop codon P-site offset (per-length)
P-site offset (per-length)
Annotate reads
Annotated read alignments (all lengths)
Frame distribution (per-length)
UTR/CDS alignment stats (per-length)
Decide read lengths to include as RPFs
Filter annotated reads by size (Ribo-seq and RNA-seq separately)
Annotated read alignments (size selected, Ribo-seq and RNA-seq separately)
mRNA coverage
Gene of Interest codon coverage
RPKM
Global codon patterns

Figure 1. Flowchart of ribosome profiling data analysis pipeline. Colored arrows indicate steps that first require execution of utility script (blue and yellow) or require manual input by the user (red). Pipeline steps are represented as ovals (main step) or pentagons (validation / analysis step). Rectangles represent input / output data. UTR: untranslated region, CDS: coding sequence, RPF: ribosome protected fragments, RPKM: reads per kilobase of transcript per million mapped reads.

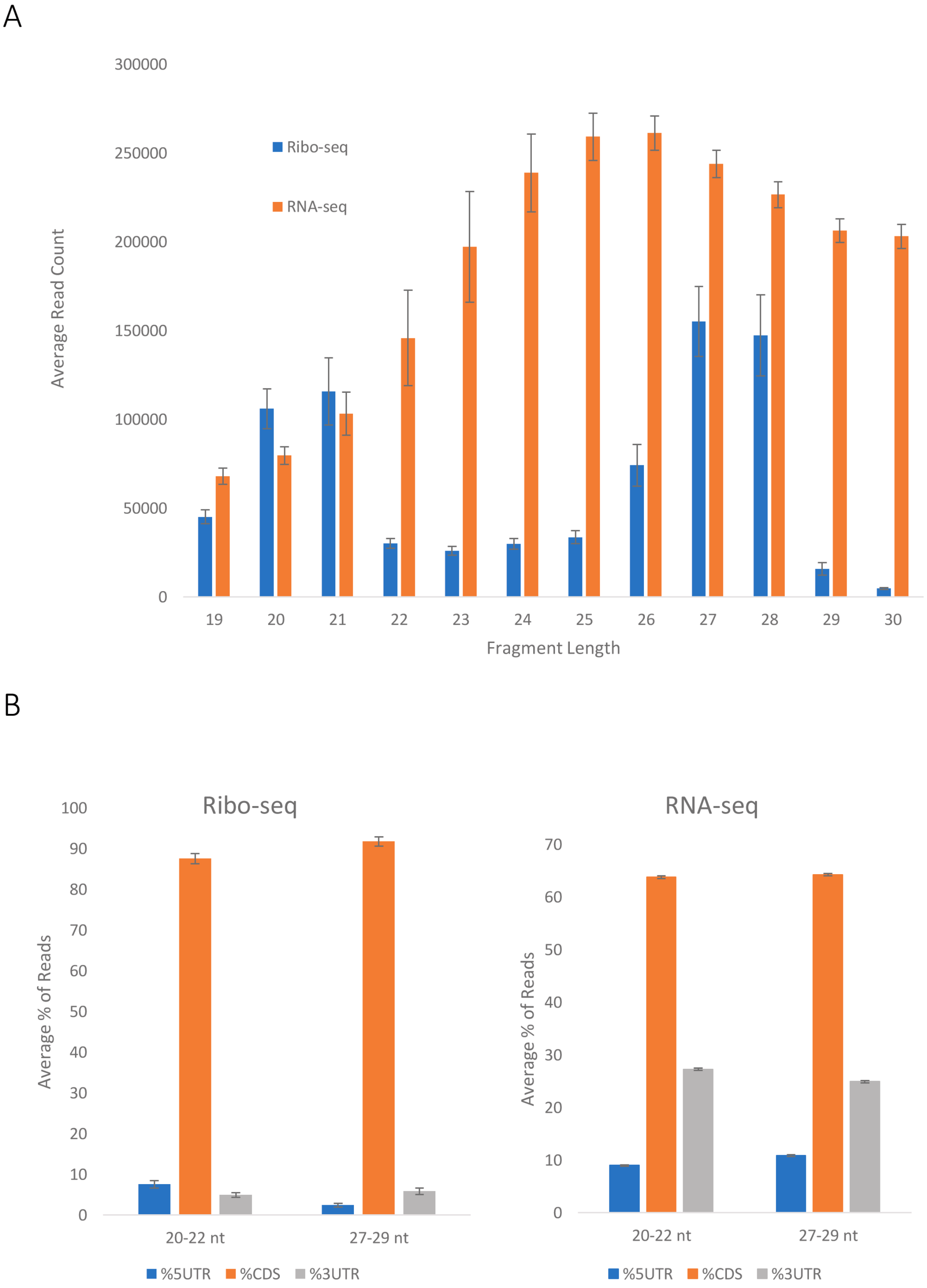

Figure 2. Ribo-Seq and RNA-Seq data distribution. (**a**) Fragment size distribution of Ribo-seq and RNA-seq reads. The average of 6 experiments (3 WT and 3 CO *F9*) was plotted, s.e.m. are shown. (**b**) Distribution of Ribo-seq (left) and RNA-seq (right) reads in mRNA coding regions (CDSs) and untranslated (5′UTR and 3′UTR) regions. The average of 6 experiments (3 WT and 3 CO *F9*) was plotted, s.e.m. are shown.

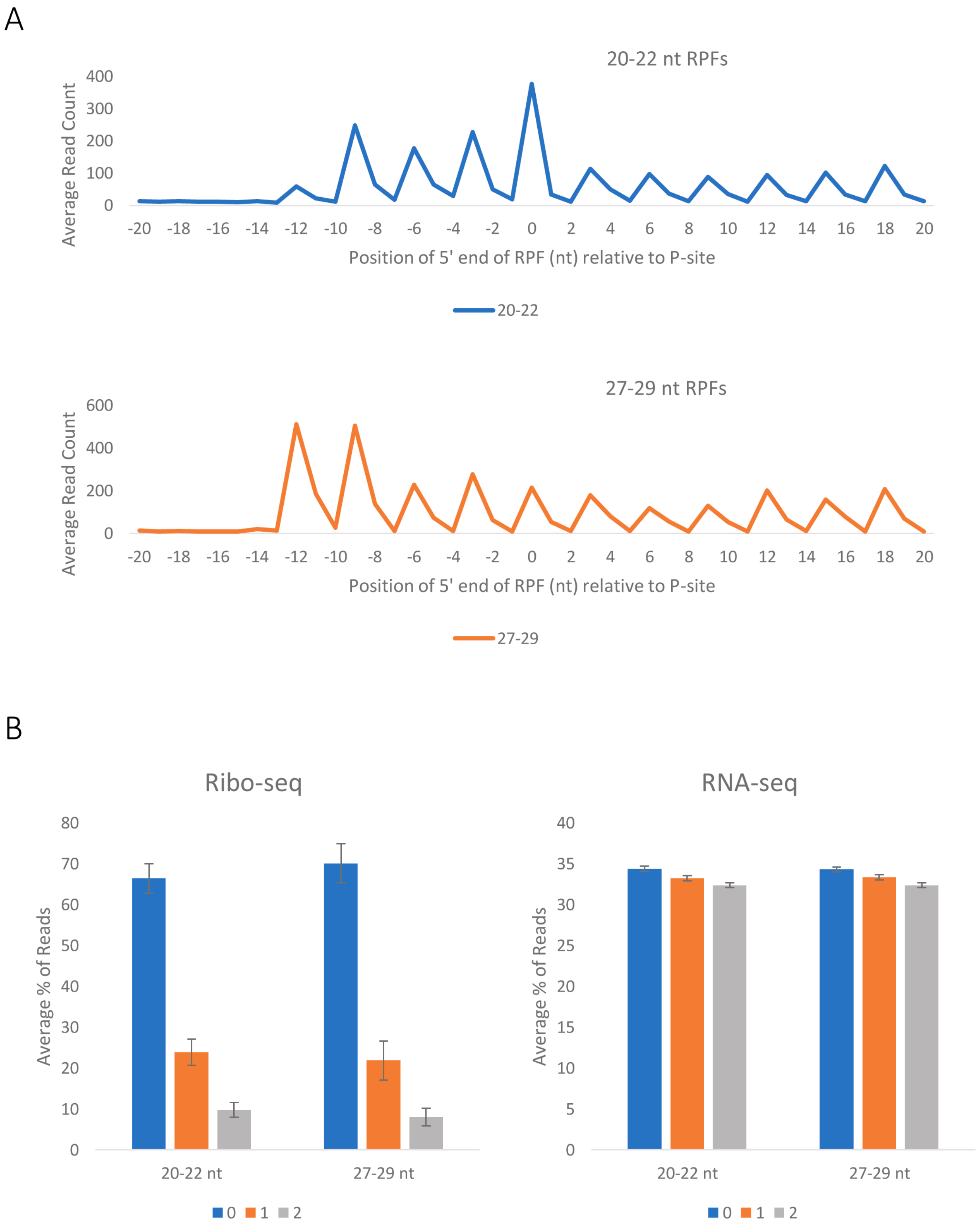

Figure 3. Triplet periodicity of Ribo-Seq data. (**a**) Profiles of the 5′ end positions of all 20–22 nt (top) and 27–29 nt (bottom) fragments relative to the start codon of their genes. The average of 6 experiments (3 WT and 3 CO *F9*) was plotted. (**b**) Positions of 20–22 nt and 27–29 nt fragments relative to the reading frame of the Ribo-seq (left) and RNA-seq (right) reads. The average of 6 experiments (3 WT and 3 CO *F9*) was plotted, s.e.m. are shown.

Table 1. Pearson and Spearman correlations between pairs of experiments. RPKM of each gene in the Ribo-seq and RNA-seq datasets were calculated, considering reads with the ribosome A site annotated at least 20 nt downstream of the start codon. A comparison between each pair of experiments within the 3 replicates was performed

Experiment	Ribo-Seq (Pearson)	RNA-Seq (Pearson)	Ribo-Seq (Spearman)	RNA-Seq (Spearman)
WT1-WT2	0.9973	0.9958	0.9426	0.9775
WT2-WT3	0.9972	0.9976	0.9513	0.9785
WT1-WT3	0.9962	0.9917	0.9384	0.9774
CO1-CO2	0.9908	0.9979	0.9314	0.9755
CO2-CO3	0.9927	0.998	0.9282	0.9771
CO1-CO3	0.994	0.9979	0.9428	0.9771

Table 2. Quality data of sequencing files. Sample ID, index, yield, number of clusters, percent Q30 and above and mean Q score for all sequencing experiments.

Sample	Index	Yield (Mbp)	#Cluster	%Q30	Mean Q
1R	CAGATC	1,669	13,355,848	66.82	25.8
1T	ATCACG	1,821	14,566,314	79.59	30.33
2R	ACTTGA	1,681	13,451,867	71.56	27.47
2T	CGATGT	1,867	14,932,652	78.55	29.96
3R	GATCAG	1,512	12,092,292	71.18	27.36
3T	TTAGGC	1,653	13,227,113	79.74	30.37
4R	TAGCTT	1,825	14,600,572	68.43	26.38
4T	TGACCA	1,731	13,848,340	79.75	30.38
5R	GGCTAC	1,537	12,292,279	67.63	26.08
5T	ACAGTG	1,754	14,033,677	80.22	30.55
6R	CTTGTA	1,818	14,541,142	68.64	26.44
6T	GCCAAT	1,662	13,296,276	78.6	29.97

USA), Bowtie (1.0.0) (http://bowtie-bio.sourceforge.net/index.shtml) (Johns Hopkins University, Baltimore, MD, USA), HISAT2 (2.1.0) (https://ccb.jhu.edu/software/hisat2/manual.shtml) (Johns Hopkins University, Baltimore, MD, USA), FASTX-Toolkit (0.0.14) (http://hannonlab.cshl.edu/fastx_toolkit/commandline.html) (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA), Samtools (1.7 using htslib 1.7) (http://www.htslib.org/) (Genome Research Limited, Hinxton, Cambridgeshire, UK).

Table 3. Description of data deposited in SRA. Filenames, SRA experiment accession, SRA sample accession and brief description of the 12 Ribo-seq and RNA-seq FASTQ files. All data files are accessible from SRA BioProject accession PRJNA591214. Data files represent three replicates of each condition (WT *F9* Ribo-seq, WT *F9* RNA-seq, CO *F9* Ribo-seq and CO *F9* RNA-seq).

Filename	SRA Experiment	SRA Sample	Description
1R_CAGATC_L002_R1_001.fastq.gz	SRX7201733	SAMN13354200	WT F9 RIBO-SEQ 1
1T_ATCACG_L002_R1_001.fastq.gz	SRX7201734	SAMN13354201	WT F9 mRNA-SEQ 1
2R_ACTTGA_L002_R1_001.fastq.gz	SRX7201737	SAMN13354202	WT F9 RIBO-SEQ 2
2T_CGATGT_L002_R1_001.fastq.gz	SRX7201738	SAMN13354203	WT F9 mRNA-SEQ 2
3R_GATCAG_L002_R1_001.fastq.gz	SRX7201739	SAMN13354204	WT F9 RIBO-SEQ 3
3T_TTAGGC_L002_R1_001.fastq.gz	SRX7201740	SAMN13354205	WT F9 mRNA-SEQ 3
4R_TAGCTT_L002_R1_001.fastq.gz	SRX7201741	SAMN13354206	CO F9 RIBO-SEQ 1
4T_TGACCA_L002_R1_001.fastq.gz	SRX7201742	SAMN13354207	CO F9 mRNA-SEQ 1
5R_GGCTAC_L002_R1_001.fastq.gz	SRX7201743	SAMN13354208	CO F9 RIBO-SEQ 2
5T_ACAGTG_L002_R1_001.fastq.gz	SRX7201744	SAMN13354209	CO F9 mRNA-SEQ 2
6R_CTTGTA_L002_R1_001.fastq.gz	SRX7201735	SAMN13354210	CO F9 RIBO-SEQ 3
6T_GCCAAT_L002_R1_001.fastq.gz	SRX7201736	SAMN13354211	CO F9 mRNA-SEQ 3

Data availability

NCBI BioProject: Ribosome profiling of HEK-293T cells stably expressing wild-type and codon-optimized coagulation factor IX. Accession number PRJNA591214; https://identifiers.org/NCBI/bioproject:PRJNA591214.

This project collates the raw data, held at the NCBI Sequence Read Archive (SRA).

Software availability

The pipeline, including the code used to process the presented dataset and instructions for use, is available: https://github.com/FDA/Ribosome-Profiling

Archived pipeline at time of publication: https://doi.org/10.5281/zenodo.3678709[14].

License: MIT License.

Acknowledgements

The authors would like to thank Dr. Nicholas T. Ingolia and Dr. Estelle Russek-Cohen, CDER, FDA for very useful discussions. This work was supported by funds from the U.S. Food and Drug Administration Chief Scientist grant and in part by an appointment to the Research Participation Program at the Center for Biologics Evaluation and Research administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the U.S. Food and Drug Administration (C.K.-S.). Additionally, this work was in part supported by the National Institutes of Health grant HL151392 (A.A.K.). This research was supported by the Intramural Research Program of the National Library of Medicine at the NIH (M.D.).

References

1. Ingolia NT, Ghaemmaghami S, Newman JR, *et al.*: **Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling.** *Science.* 2009; **324**(5924): 218–223.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

2. Woolstenhulme CJ, Guydosh NR, Green R, *et al.*: **High-precision analysis of translational pausing by ribosome profiling in bacteria lacking EFP.** *Cell Rep.* 2015; **11**(1): 13–21.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

3. Gobet C, Naef F: **Ribosome profiling and dynamic regulation of translation in mammals.** *Curr Opin Genet Dev.* 2017; **43**: 120–127.
PubMed Abstract | **Publisher Full Text**

4. Ingolia NT: **Ribosome Footprint Profiling of Translation throughout the**

Genome. *Cell.* 2016; **165**(1): 22–33.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Alexaki A, Hettiarachchi GK, Athey JC, *et al.*: **Effects of codon optimization on coagulation factor IX translation and structure: Implications for protein and gene therapies.** *Sci Rep.* 2019; **9**(1): 15449.
PubMed Abstract | Publisher Full Text | Free Full Text

6. Hunt, R, Hettiarachchi G, Katneni U, *et al.*: **A Single Synonymous Variant (c.354G>A [p.P118P]) in *ADAMTS13* Confers Enhanced Specific Activity.** *Int J Mol Sci.* 2019; **20**(22): 5734.
PubMed Abstract | Publisher Full Text | Free Full Text

7. Hettiarachchi GK, Katneni UK, Hunt RC, *et al.*: **Translational and transcriptional responses in human primary hepatocytes under hypoxia.** *Am J Physiol Gastrointest Liver Physiol.* 2019; **316**(6): G720–G734.
PubMed Abstract | Publisher Full Text | Free Full Text

8. Fields AP, Rodriguez EH, Jovanovic M, *et al.*: **A Regression-Based Analysis of Ribosome-Profiling Data Reveals a Conserved Complexity to Mammalian Translation.** *Mol Cell.* 2015; **60**(5): 816–827.
PubMed Abstract | Publisher Full Text | Free Full Text

9. Ingolia NT, Lareau LF, Weissman JS: **Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes.** *Cell.* 2011; **147**(4): 789–802.
PubMed Abstract | Publisher Full Text | Free Full Text

10. Young DJ, Guydosh NR, Zhang F, *et al.*: **Rli1/ABCE1 Recycles Terminating Ribosomes and Controls Translation Reinitiation in 3′UTRs *In Vivo*.** *Cell.* 2015; **162**(4): 872–884.
PubMed Abstract | Publisher Full Text | Free Full Text

11. Spealman P, Naik AW, May GE, *et al.*: **Conserved non-AUG uORFs Revealed by a Novel Regression Analysis of Ribosome Profiling Data.** *Genome Res.* 2018; **28**(2): 214–222.
PubMed Abstract | Publisher Full Text | Free Full Text

12. Michel AM, Ahern AM, Donohue CA, *et al.*: **GWIPS-viz as a Tool for Exploring Ribosome Profiling Evidence Supporting the Synthesis of Alternative Proteoforms.** *Proteomics.* 2015; **15**(14): 2410–2416.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Tunney R, McGlincy N, Graham ME, *et al.*: **Accurate Design of Translational Output by a Neural Network Model of Ribosome Distribution.** *Nat Struct Mol Biol.* 2018; **25**(7): 577–582.
PubMed Abstract | Publisher Full Text | Free Full Text

14. CBER-Source: **FDA/Ribosome-Profiling: Ribosome profiling (Version v1.0).** *Zenodo.* 2020.
http://www.doi.org/10.5281/zenodo.3678709

15. Suwanmanee T, Hu G, Gui T, *et al.*: **Integration-deficient lentiviral vectors expressing codon-optimized R338L human FIX restore normal hemostasis in Hemophilia B mice.** *Mol Ther.* 2014; **22**(3): 567–574.
PubMed Abstract | Publisher Full Text | Free Full Text

16. Lareau LF, Hite DH, Hogan GJ, *et al.*: **Distinct Stages of the Translation Elongation Cycle Revealed by Sequencing Ribosome-Protected mRNA Fragments.** *eLife.* 2014; **3**: e01257.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Peer Review Status: ✓ ✓ ✓

Version 2

Reviewer Report 07 October 2020

https://doi.org/10.5256/f1000research.29425.r71732

✓

James Collawn
Department of Cell, Developmental and Integrative Biology, University of Alabama at Birmingham, Birmingham, AL, USA

I am fine with the responses of the authors. I have no further concerns.

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 28 September 2020

https://doi.org/10.5256/f1000research.29425.r71730

✓

Stefano Biffo
INGM, National Institute of Molecular Genetics, “Fondazione Romeo ed Enrica Invernizzi”, Milan, Italy

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 22 September 2020

https://doi.org/10.5256/f1000research.29425.r71731

✓

Jordan Berg
Department of Biochemistry, University of Utah, Salt Lake City, UT, USA

The changes look good. My only final recommendation would be to version the pipeline again with the updates and make sure the updated repository is archived in Zenodo (it looks like it only has the version from Feb 2020 archived).

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Ribosome profiling library creation and data analysis.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 14 July 2020

https://doi.org/10.5256/f1000research.24715.r65798

?

Rafal Bartoszewski
Department of Biology and Pharmaceutical Botany, Medical University of Gdansk, Gdansk, Poland
James Collawn
Department of Cell, Developmental and Integrative Biology, University of Alabama at Birmingham, Birmingham, AL, USA

This study by Dr. Kimchi-Sarfaty examines ribosome profiling to follow the translational kinetics of factor IX (F9) and is a follow up of a previous study of theirs (DOI: 10.1038/s41598-019-51984-2)[1]. They compare two HEK293T cell lines with one expressing a WT F9 gene and another with a codon-optimized F9 gene. In their previous study, they demonstrated the importance of this type of analysis by illustrating that these two cell lines had different translational kinetics and furthermore, different effects on protein conformation. In this study, they compared 3 replicates of the RNA-seq and Ribo-seq of the two cell lines. The correlations between the pairs of experiments seemed quite good and we believe this is an important type of study, however, we do have two concerns. First of all, the housekeeping control genes they utilize are GAPDH and ACTB, both of which are translated on free ribosomes in the cytosol. F9, however, is a secreted protein that is translated on ER ribosomes. Therefore, a rationale for the appropriateness of the two controls needs to be discussed. Secondly, they used a Box-Cox variance-stabilizing transformation for raw data followed by a Kolmogorov-Smirnov test which tests for probability distribution functions but is a less selective normality test in our view. Why not utilize Saphiro-Wilk (https://doi.org/10.1093/biomet/52.3-4.591)[ref-2] or Arednson-Darling tests (doi:10.2307/2281537)[3]? Some more detail here about the rationale for the statistical analyses is therefore warranted and whether additional controls and samples would have any effect on the outcomes of the studies should be discussed.

References

1. Alexaki A, Hettiarachchi G, Athey J, Katneni U, et al.: Effects of codon optimization on coagulation factor IX translation and structure: Implications for protein and gene therapies. *Scientific Reports*. 2019; **9** (1). Publisher Full Text
2. SHAPIRO S, WILK M: An analysis of variance test for normality (complete samples). *Biometrika*. 1965; **52** (3-4): 591-611 Publisher Full Text
3. Anderson T, Darling D: A Test of Goodness of Fit. *Journal of the American Statistical Association*. 1954; **49** (268). Publisher Full Text

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Yes

Are sufficient details of methods and materials provided to allow replication by others?

Yes

Are the datasets clearly presented in a useable and accessible format?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Protein trafficking, protein expression and folding (Collawn); mRNA structure, translational expression, and silent polymorphisms (Bartoszewski).

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Author Response ( ) 04 Sep 2020

Chava Kimchi-Sarfaty, Food and Drug Administration, USA, Silver Spring, USA

Thank you very much for raising this interesting question. Although they are not secreted proteins as F9, ACTB and GAPDH were chosen as they have a relatively high level of expression and therefore generate sufficient sequencing reads to lead to a well resolved ribosome profile. Moreover, these two genes' expression is stable and not changing due to the transfection or cell culture growth. It is not our intention to make any conclusions regarding the translation kinetics of secreted versus non-secreted proteins. Rather, the control genes are compared to themselves between experimental groups (cells expressing WT F9 and cells expressing CO F9). In this context, we believe that ACTB and GAPDH are appropriate to use as control genes.

Regarding the statistical method, we want to point out that the Kolmogorov-Smirnov method is considered more conservative than the methods mentioned by the reviewer. Therefore, we feel that by using it we are able to detect a difference between two cumulative curves, then we should be more confident that the difference is real. Second, our aim is not to test for normality, but rather to compare two distributions. The main aims of the normalization step via the Box-Cox transformation is to stabilize the variance (of the originally skewed distribution), and to put the distributions on a comparable scale, so that we can compare the cumulative plots.

Competing Interests: No competing interests were disclosed.

Reviewer Report 13 July 2020

https://doi.org/10.5256/f1000research.24715.r65801

?

Jordan Berg

Department of Biochemistry, University of Utah, Salt Lake City, UT, USA

The provided dataset provides the opportunity for more in-depth study of two versions of the *F9* gene with identical protein amino acid sequence but different nucleotide coding sequences, thus allowing for an exploration of the consequences of differential codon usage. The dataset itself is of biological and pharmacological interest and offers a valuable data contribution. More specific comments and questions regarding key points of the data note are provided below.

Introduction:

- At face value, this seems like a re-publishing of a dataset that has already been described in the literature (albeit with some additional detail and the pipeline). Were any data added or modified between the original publication and this data note? It is not clear why this dataset was not published with the original study describing these data (https://doi.org/10.1038/s41598-019-51984-2)[1].

- If part of the intended contribution of this data note is the pipeline, it would be helpful to at least provide a bash script that ties together the different scripts and flexibly accepts input FASTQ files. Currently, with all the hard-coding present, it makes it difficult to reuse these scripts, especially when input references or sequence files vary from those used when analyzing this particular dataset.

Methods:

- A Pearson's correlation assumes the data are normally distributed. As sequencing data follows a negative binomial distribution, the use of a Spearman rank-order coefficient would be more appropriate. It might be helpful to clarify that this is for comparing biological replicates.

- Is there a reason the A-site offset is set at a strict 15 nt? Recent methods, such as those found in the RiboWaltz (https://doi.org/10.1371/journal.pcbi.1006169)[2] package allow for a more optimized P-site offset determination that could probably be applied to A-site offset determination. On that note, is there a downstream reason you are interested in calculating the offset of the A-site? It looks like P-sites were used when calculating periodicity in the figures and scripts and thus it is not clear why the A-site offset is mentioned.

- The pipeline uses deprecated software (TopHat has been deprecated for nearly 5 years now) and should be updated to use a more accurate splice-aware option (such as STAR, HISAT2). Figures might need to be updated as appropriate. Is the installation of dependencies included in a script, or does the user need to handle that? For example, BioPython and pysam are included as a dependency in the python scripts, but I don't see them listed in the manuscript.

GitHub Documentation:

- To aid in readability, some formatting updating (new lines in example code) would be helpful. Currently, there are several commands on the same line as a comment, making the commands and comments difficult to read. Using markdown syntax to display the code sections would aid in readability as well.

- Explicitly state the gene annotation used for the preparation of this dataset. Comprehensive, basic gene, etc? Same for the GFF3 file. It is not clear from the documentation itself what versions were used and would be nice from an archiving/reproducibility stand-point. The same goes for listing software versions used for processing the dataset as presented in the paper.

Code:

- I could not get the *trim-adapters.py* script to work using the deposited data. I had empty trimmed files output and no error information was displayed to aid in debugging.

- As far as functionality, all of the indexing scripts appeared to function, but I was unable to test past the trim-adapters.py script due to the issue mentioned above.

References

1. Alexaki A, Hettiarachchi G, Athey J, Katneni U, et al.: Effects of codon optimization on coagulation factor IX translation and structure: Implications for protein and gene therapies. *Scientific Reports*. 2019; **9** (1). Publisher Full Text
2. Lauria F, Tebaldi T, Bernabò P, Groen EJN, et al.: riboWaltz: Optimization of ribosome P-site positioning in ribosome profiling data.*PLoS Comput Biol*. **14** (8): e1006169 PubMed Abstract | Publisher Full Text

Is the rationale for creating the dataset(s) clearly described?
Yes

Are the protocols appropriate and is the work technically sound?
Partly

Are sufficient details of methods and materials provided to allow replication by others?
Partly

Are the datasets clearly presented in a useable and accessible format?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Ribosome profiling library creation and data analysis.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response ( ) 04 Sep 2020

Chava Kimchi-Sarfaty, Food and Drug Administration, USA, Silver Spring, USA

Thank you very much for your thorough review. Below we provide a point by point response.

- At face value, this seems like a re-publishing of a dataset that has already been described in the literature (albeit with some additional detail and the pipeline). Were any data added or modified between the original publication and this data note? It is not clear why this dataset was not published with the original study describing these data (https://doi.org/10.1038/s41598-019-51984-2)[1].

Regrettably, due to space restrictions we did not have the opportunity to publish the pipeline and the raw data with a detailed description in the original publication, which focused on the comparison of the wild type and codon optimized coagulation factor 9. We strongly believe that the raw ribosome profiling data can be of value to the scientific community and should be public with sufficient description to make it accessible. Similarly, we hope that the pipeline could be of use to researchers, and we note that despite the multitude of ribosome profiling papers that have been published so far, the software used in the analysis is not always readily accessible.

- If part of the intended contribution of this data note is the pipeline, it would be helpful to at least provide a bash script that ties together the different scripts and flexibly accepts input FASTQ files. Currently, with all the hard-coding present, it makes it difficult to reuse these scripts, especially when input references or sequence files vary from those used when analyzing this particular dataset.

We appreciate the suggestion and have made changes accordingly. The pipeline has been modified to be more user friendly, and updates have been made as described in the response to Reviewer 1.

Methods:

- A Pearson's correlation assumes the data are normally distributed. As sequencing data follows a negative binomial distribution, the use of a Spearman rank-order coefficient would be more appropriate. It might be helpful to clarify that this is for comparing biological replicates.

Thank you for pointing this out. We have added a Spearman rank-order coefficient to Table 1 for each of the dataset comparisons. The text describing Table 1 has been updated.

- Is there a reason the A-site offset is set at a strict 15 nt? Recent methods, such as those found in the RiboWaltz (https://doi.org/10.1371/journal.pcbi.1006169)[2] package allow for a more optimized P-site offset determination that could probably be applied to A-site offset determination. On that note, is there a downstream reason you are interested in calculating the offset of the A-site? It looks like P-sites were used when calculating periodicity in the figures and scripts and thus it is not clear why the A-site offset is mentioned.

Thank you for pointing out the lack of clarity regarding the A- and P-site offset. The A-site offset at 15 nt is equivalent to a P-site offset at 12 nt. We made changes to consistently refer to the P-site offset. We agree that when the translation frame is undetermined a hard 12 nt P-site offset can lead to inaccuracies, and an optimized method for P-site determination is paramount. However, our analysis did not incorporate unannotated open reading frames (ORFs), non-conventional translation initiation sites and we did not attempt to reveal novel translated regions. As a result, the triplet periodicity in our dataset is unambiguous and the P-site offset can be determined at 12 nt. For ribosome-protected fragments that aligned in reading frame one or two, a +/- 1 offset was additionally added to properly annotate the codons within each site of the ribosome.

- The pipeline uses deprecated software (TopHat has been deprecated for nearly 5 years now) and should be updated to use a more accurate splice-aware option (such as STAR, HISAT2). Figures might need to be updated as appropriate. Is the installation of dependencies included in a script, or does the user need to handle that? For example, BioPython and pysam are included as a dependency in the python scripts, but I don't see them listed in the manuscript.

We agree with your concern. When we started developing our pipeline, TopHat was the software of choice and we retained it to allow duplication of the data in our published paper [1]. We have now updated the pipeline to use HISAT2; all the figures have been updated accordingly. The readme file and manuscript have been updated to reflect the complete list of software dependencies and their versions.

GitHub Documentation:

- To aid in readability, some formatting updating (new lines in example code) would be helpful. Currently, there are several commands on the same line as a comment, making the commands and comments difficult to read. Using markdown syntax to display the code sections would aid in readability as well.

We thank you very much for pointing this out. We have removed all unnecessary commands and lines of code that had been commented out. We believe this will aid in readability of the code.

- Explicitly state the gene annotation used for the preparation of this dataset. Comprehensive, basic gene, etc? Same for the GFF3 file. It is not clear from the documentation itself what versions were used and would be nice from an archiving/reproducibility stand-point. The same goes for listing software versions used for processing the dataset as presented in the paper.

We agree with your concern. We have updated the GitHub documentation to explicitly state which annotation and assembly versions were used in the in the analysis of this dataset.

- I could not get the *trim-adapters.py* script to work using the deposited data. I had empty trimmed files output and no error information was displayed to aid in debugging.As far as functionality, all of the indexing scripts appeared to function, but I was unable to test past the trim-adapters.py script due to the issue mentioned above.

Thank you very much for pointing this out. This is an interesting result. We have re-run the analysis pipeline using the new bash scripts and code updated to Python 3.7, and the adapter trimming step ran correctly. However, we have added an error message in the Trim_adapters.py script to notify the user if an error occurs to aid in the debugging process.

- Are the protocols appropriate and is the work technically sound? Partly

We have updated the alignment tool used in the pipeline to HISAT2 and the code used in the individual scripts to Python 3.7. We hope that these updates to the pipeline have improved the appropriateness of the protocol.

- Are sufficient details of methods and materials provided to allow replication by others? Partly

We have bundled all scripts in the pipeline into two shell scripts and there is now a warning message in the Trim_adapters.py script that notifies the user when it does not execute properly. We have also updated the dependency list to accurately reflect all additional software required as well as versions that were tested. We hope that these updates have made replication of the work easier for the user.

- Are the datasets clearly presented in a useable and accessible format? Partly

We hope that the updates to the pipeline described above have made the datasets more useable and accessible.

1 Alexaki, A. *et al.* Effects of codon optimization on coagulation factor IX translation and structure: Implications for protein and gene therapies. *Sci Rep* **9**, 15449, (2019).

Competing Interests: No competing interests were disclosed.

Reviewer Report 05 June 2020

https://doi.org/10.5256/f1000research.24715.r64172

?

Stefano Biffo
INGM, National Institute of Molecular Genetics, “Fondazione Romeo ed Enrica Invernizzi”, Milan, Italy
Riccardo Rossi
INGM, Milan, Italy

The report is certainly interesting and the possibility to access data relatively nice. However, in the present form there is not much advantage compared to a paper supplementary section and accession number.
In the current pipeline one should install - as per explicit requirements - Python 2.7, Bowtie, Tophat, Samtools, which still requires bioinformatician work.

Some programs are not used anymore. Python 2 has been disconnected since January. Tophat as mapping resource is not suggested even by the developers, since now there are better ones. In the present form the pipeline can clearly work. In order to facilitate the use of the pipeline the authors could have wrapped everything in a folder that could have been run by the readers more easily.
The biological data seem very nice to me (Stefano Biffo with the help of a bioinformatician).

Is the rationale for creating the dataset(s) clearly described?
Yes

Are the protocols appropriate and is the work technically sound?
Yes

Are sufficient details of methods and materials provided to allow replication by others?
Yes

Are the datasets clearly presented in a useable and accessible format?
No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Translational control of gene expression

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Author Response ( ) 19 Jun 2020

Chava Kimchi-Sarfaty, Food and Drug Administration, USA, Silver Spring, USA

Dr. Biffo,

Thank you very much for the time you put into reviewing the article and for your suggestions. We would like to inform you that we are currently working on upgrading the pipeline as per your comments. We will be updating the code to Python 3, bundling individual Python scripts into a more easily executable shell script and replacing the current Tophat alignment with a more modern tool. We very much appreciate your comments and we will notify you when we have completed the upgrades.

Chava Kimchi-Sarfaty and Jacob Kames

Competing Interests: No competing interests were disclosed.

Author Response ( ) 04 Sep 2020

Chava Kimchi-Sarfaty, Food and Drug Administration, USA, Silver Spring, USA

We appreciate the suggestion and have made changes accordingly. The Python scripts used in the code have been updated to Python 3.7. Furthermore, the alignment software used in the pipeline has been changed from TopHat to Hisat2. Finally, two bash scripts have been created to run the pipeline. The first, *build_hisat_index.sh*, carries out the commands to set up the reference index with the two F9 constructs used in this study. The second, *RP_analysis_pipeline.sh*, runs the remaining steps of the pipeline. Both are run from within the *Ribosome_profiling* directory and have the dataset defined within each script.

We hope the reviewer agrees that the current changes make the software more user friendly and thus usable and accessible.

Competing Interests: No competing interests were disclosed.