Fivepseq: analysis of 5′ count distribution in RNA-seq datasets for inference on mRNA degradation and translation

Nersisyan Lilit[1], Ropat Maria[1] and Pelechano Vicent[1*]

[1]SciLifeLab, Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Solna, 171 65, Sweden

*To whom correspondence should be addressed.

Abstract

Motivation: The emergence of techniques for the analysis of 5'-monophosphate mRNA degradation intermediates necessitates development of tools for automated analysis of 5' endpoint distribution for inference on ribosome dynamics, mRNA cleavage patterns, binding sites, etc.

Results: Here we present fivepseq: an easy-to-use command-line application for analysis and visualization of 5' endpoint count distribution from RNA-seq datasets. It produces interactive reports for ease of exploratory analysis that provide single-nucleotide resolution reports on ribosome dynamics and degradation patterns.

Availability: Fivepseq is available from http://pelechanolab.com/software/fivepseq, under BSD 3-Clause License.

Contact: vicente.pelechano.garcia@ki.se

Introduction

The functional status of living cells largely depends on regulation of the pool of translating mRNAs, which are degraded by the action of exo- and endo-nucleases, producing diverse degradation intermediates. 5' monophosphorylated (5P) intermediates can be captured and sequenced by techniques such as 5PSeq (1), PARE (2) or GMUC (3). It has recently been shown that at least in yeast [1] and in plants (2), mRNA degradation can occur co-translationally. This allows for using 5P mRNA degradation intermediates as a fast readout for the position of the last translating ribosome (1–5). Consequently, novel bioinformatics pipelines are needed in order to (i) make inferences on translational states from 5P count distribution by exploring translational frame preference and presence of 3-nucleotide (nt) periodicity; (ii) explore ribosome pausing/stalling at initiation/termination via analysis of 5P count distribution relative to translation start/stop sites; (iii) determine slow/fast codons by mapping genome-wide 5P distribution relative to codons/amino acids. Currently available software solutions are

mostly designed for ribosome profiling datasets and provide only partial functionalities for abovementioned analyses; some require bioinformatics expertise, while others implement server based solutions that are not scalable to high volume data (6,7). In addition, they often expect parameters such as “ribosome protection size” that are not relevant for in vivo 5’-3’ co-translational degradation.

Here we present fivepseq, a standalone command line application that performs comprehensive analyses of 5’ endpoint distribution from any RNA-seq data. The package is easy to use, and comes with a wide range of visualizations in the form of interactive reports that allow for smooth exploratory analyses and knowledge inference.

Methods

The fivepseq package is written in python 2.7 and can be used with python 2.7 or 3.x in Unix operating systems. It relies on the plastid package for parsing of alignment files (8), and on the bokeh package (9) for visualization.

Fivepseq parses alignment BAM and genome annotation files to derive counts of read 5’ endpoints at each genomic position in protein coding genes. Noise removal is performed by scaling down counts falling outside of Poisson distribution with sample-specific mean values. Library size normalization is performed per million of mapped reads in the coding regions. Finally, the user can also limit the analysis to a specified set of genes under interest.

The primary output of fivepseq are HTML formatted interactive reports that contain plots describing the distribution of 5’ endpoints relative to translation start and end sites or codons and translational frame preference at genome-wide and gene-level scales. With additional options, these patterns can be compared across different gene sets provided by the user. In addition, a variety of plain text files containing raw counts are exported in convenient formats.

Datasets

To show its wide applicability, we have used 5PSeq data from Saccharomyces cerevisiae (10) (GSE91064). We have also generated fivepseq reports for PARE datasets from Arabidopsis thaliana (GSE77549) (2), and human ribosome profiling data (GSE79664). The reports are available in the supplementary material.

Results and discussion

We have previously shown that depletion of the translation elongation factor eIF5A induces ribosome stalling at translation initiation and termination sites and at polyproline motifs (10). Here, we reproduce these results with fivepseq (Figure 1).

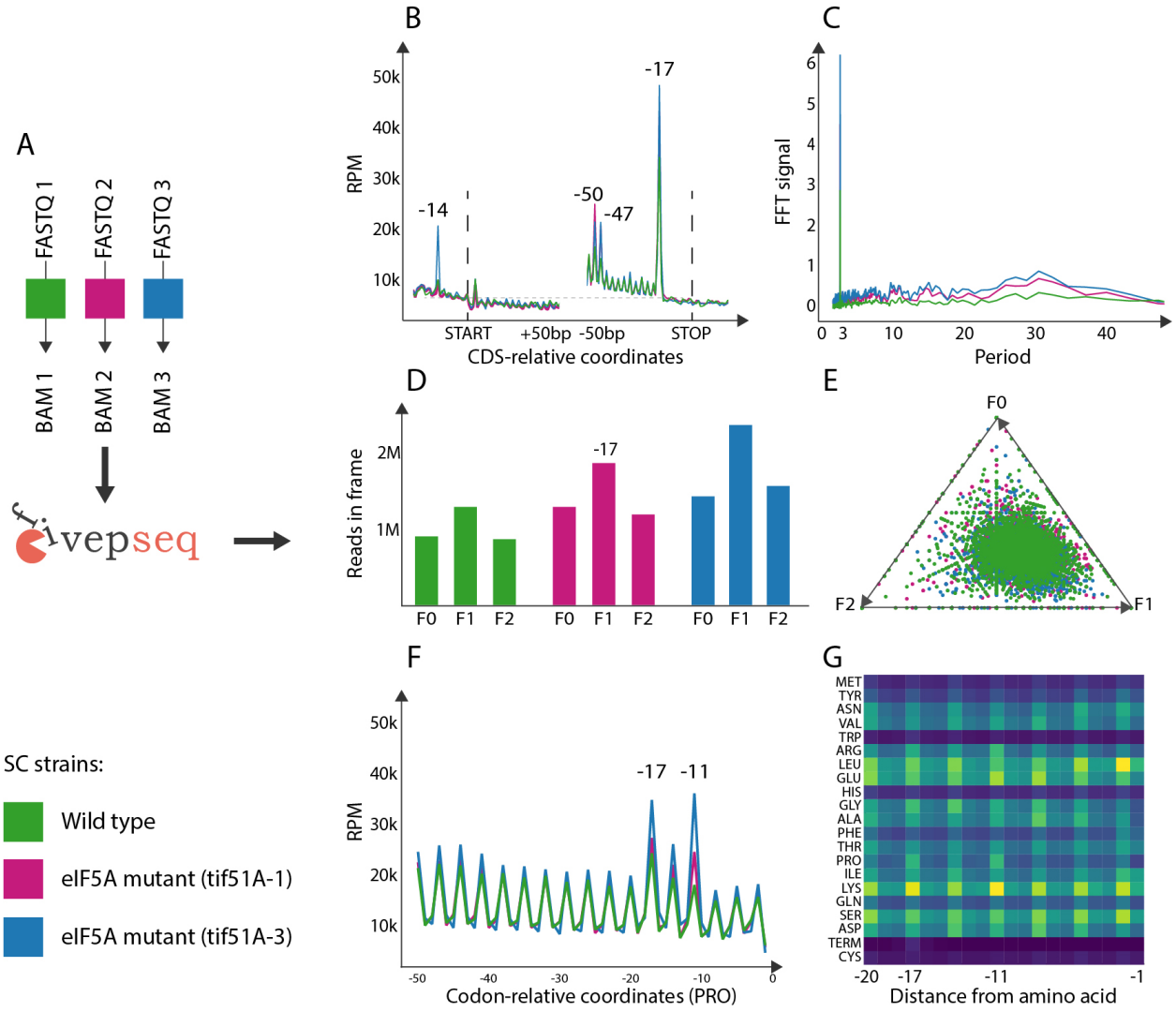

Figure 1. Fivepseq workflow. **A.** Alignment files are supplied to fivepseq. **B.** Meta-level linechart of 5' counts (reads per million, RPM) relative to translation start and stop sites. **C.** Fast Fourier transform (FFT) signals for each 5' count periodicity value. **D.** Raw 5' counts falling into frames F0, F1 or F2 at genome wide scale. **E.** Frame preference per gene (dot). **F.** Genome-wide RPM values of 5' counts relative to proline (PRO) codons. **G.** Genome-wide 5' counts at relative distances from each amino acid. Low-to-high values are color-coded on the blue-to-yellow scale.

Fivepseq aligns reads either at CDS start or stop positions and combines coordinate-wise 5' counts to derive metagene line charts (Figure 1B). These line charts show higher 5' counts every 3-nts, which is confirmed by Fast Fourier transform analysis (Figure 1C) and may indicate co-translational generation of

the majority of 5P mRNA intermediates [1]. Higher 5' counts in one or both eIF5A mutants are more prominent at positions 14nt and 17nt upstream from translation start and stop sites, respectively, indicative of ribosome stalling at these sites.

Preferable accumulation of 5' counts in the middle translation frame is observed at genome-wide level (Figure 1D) and at single-gene level, as seen in triangle plots that convert gene-specific frame counts into 2D coordinate space (Figure 1E).

The 5' count peak at 17nt upstream from proline (PRO) indicates on ribosome stalling at this amino acid (Figure 1F), while the additional peak at -11nt agrees with ribosome pausing at poly-proline motifs as previously shown (10). 5' count distribution relative to the other amino-acids may also be seen with heatmaps (Figure 1G).

Fivepseq provides options for either filtering the genes to a list under interest (or a single one), or to compare 5' count distribution patterns between various gene sets, provided by the user in a text file (see Supplementary material).

Importantly, the HTML formatted reports contain navigation tools that ease exploratory analysis. The user may limit the view to selected samples, may zoom into regions of interest or simply hover over the counts to obtain precise information at single nucleotide resolution.

Acknowledgements

We thank PelechanoLab members for discussions and alpha testing, and A. Arakelyan from BIG NAS RA and H. Khachatryan, K. Hambardzumyan from YerevaNN (Armenia) and T. Shahverdyan for tech-support and math-advice. Computational resources were provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX).

Funding

This project has received funding from the Swedish Foundation's Starting Grant (Ragnar Söderberg Foundation), the Swedish Research Council [VR 2016-01842], a Wallenberg Academy Fellowship, and Karolinska Institutet (SciLifeLab Fellowship, SFO and KI funds) to V.P.; the EU H2020-MSCA-IF-2018 program under grant agreement [845495 - TERMINATOR] to L.N.

Conflict of Interest: none declared.

References

1. Pelechano V, Wei W, Steinmetz LM. Widespread co-translational RNA decay reveals ribosome dynamics. Cell. 2015;161(6):1400–12.
2. Hou CY, Lee WC, Chou HC, Chen AP, Chou SJ, Chen HM. Global analysis of truncated RNA ends reveals new insights into Ribosome Stalling in plants. Plant Cell. 2016 Oct 1;28(10):2398–416.
3. Willmann MR, Berkowitz ND, Gregory BD. Improved genome-wide mapping of uncapped and cleaved transcripts in eukaryotes-GMUCT 2.0. Vol. 67, Methods. Academic Press Inc.; 2014. p. 64–73.
4. Wei W, Hennig BP, Wang J, Zhang Y, Piazza I, Pareja Sanchez Y, et al. Chromatin-sensitive cryptic promoters putatively drive expression of alternative protein isoforms in yeast. Genome Res [Internet]. 2019 Nov 18 [cited 2019 Dec 20]; Available from: http://www.ncbi.nlm.nih.gov/pubmed/31740578
5. Mazzoni-Putman SM, Stepanova AN. A plant biologist's toolbox to study translation. Front Plant Sci. 2018 Jul 2;9.
6. Hardcastle TJ. riboSeqR [Internet]. Available from: 10.18129/B9.bioc.riboSeqR
7. Wang H, Wang Y, Xie Z. Computational resources for ribosome profiling: from database to Web server and software. Brief Bioinform [Internet]. 2019 Jan 18;20(1):144–55. Available from: https://academic.oup.com/bib/article/20/1/144/4082612
8. Dunn JG, Weissman JS. Plastid: Nucleotide-resolution analysis of next-generation sequencing and genomics data. BMC Genomics. 2016 Nov 22;17(1).
9. Team BD. Bokeh: Python library for interactive visualization [Internet]. 2019. Available from: https://bokeh.org/
10. Pelechano V, Alepuz P. eIF5A facilitates translation termination globally and promotes the elongation of many non polyproline-specific tripeptide sequences. Nucleic Acids Res [Internet]. 2017 Jul 7 [cited 2019 Oct 1];45(12):7326. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28549188