Shoelaces: an interactive tool for ribosome profiling processing and visualization

Åsmund Birkeland[1†], Katarzyna Chyżyńska[2†] and Eivind Valen[2,3*]

[1]Department of Informatics, University of Bergen, 5008 Bergen, Norway. [2]Computational Biology Unit, Department of Informatics, University of Bergen, 5008 Bergen, Norway. [3]Sars International Centre for Marine Molecular Biology, University of Bergen, 5008 Bergen, Norway. [†]Equal contributor. [*] Correspondence: eivind.valen@uib.no

Abstract

The emergence of ribosome profiling to map actively translating ribosomes has laid the foundation for a diverse range of studies on translational regulation. The data obtained with different variations of this assay is typically manually processed, which has created a need for tools that would streamline and standardize processing steps.

We present Shoelaces, a toolkit for ribosome profiling experiments automating read selection and filtering to obtain genuine translating footprints. Based on periodicity, favoring enrichment over the coding regions, it determines the read lengths corresponding to bona fide ribosome protected fragments. The specific codon under translation (P-site) is determined by automatic offset calculations resulting in sub-codon resolution. Shoelaces provides both a user-friendly graphical interface for interactive visualisation in a genome browser-like fashion and a command line interface for integration into automated pipelines. We process 79 libraries and show that studies typically discard excessive amounts of data in their manual analysis pipelines.

Shoelaces streamlines ribosome profiling analysis offering automation of the processing, a range of interactive visualization features and export of the data into standard formats. Shoelaces stores all processing steps performed in an XML file that can be used by other groups to exactly reproduce the processing of a given study. We therefore anticipate that Shoelaces can aid researchers by automating what is typically performed manually and contribute to the overall reproducibility of studies. The tool is freely distributed as a Python package, with additional instructions and demo datasets available at https://bitbucket.org/valenlab/shoelaces.

Background

Ribosome profiling provides the first opportunity to monitor the behavior of translating ribosomes on a transcriptome-wide scale. Since its development [1], the technique has been widely adopted and inspired a diverse range of studies on translational regulation. While the assay itself has been partially standardized, the processing of data has not. A significant bottleneck is that of reproducibility and interpretation. In particular, most studies relies on manual selection of read lengths and manual P-site determination. The choices made are highly variable between studies and make it challenging to compare results in the literature.

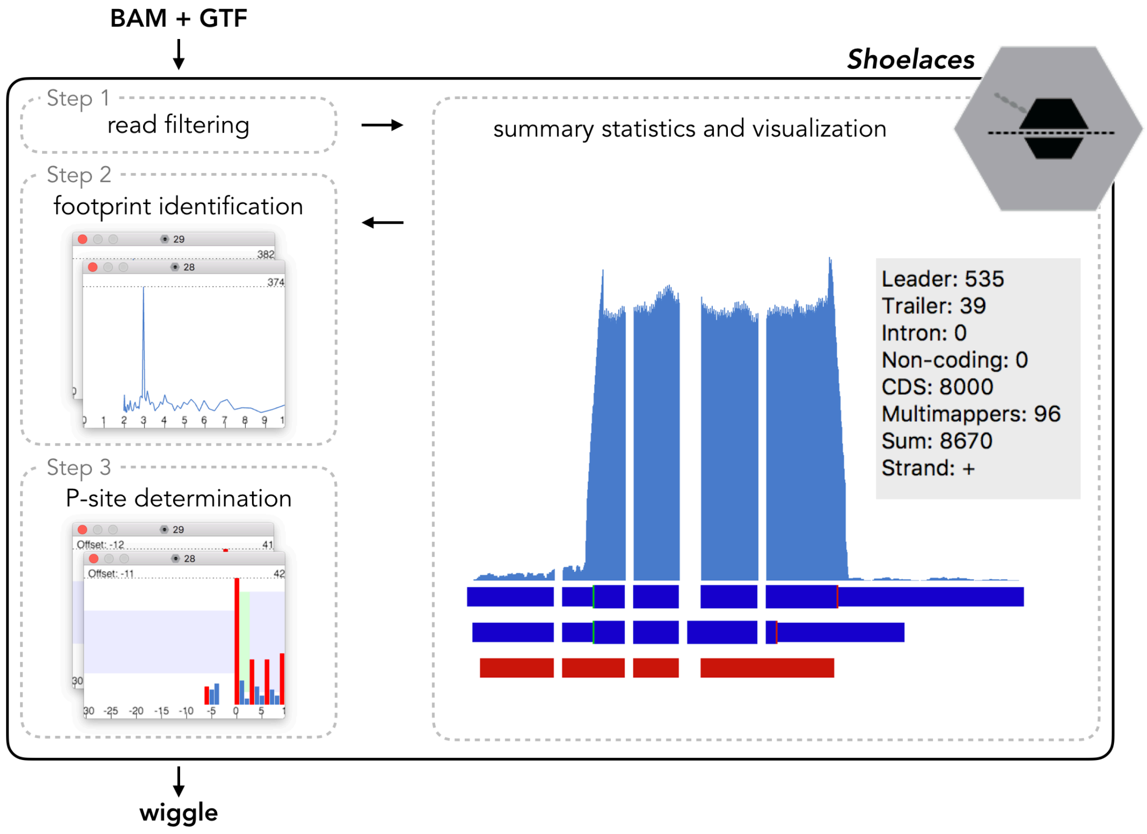
The consistent processing of such data necessitates that two major challenges are met: (1) separating signal from noise, i.e. distinguishing footprints of translating ribosomes from reads originating from other processes and (2) determining the specific codon being translated by the ribosome which the read fragment originates from (a P-site offset). While some software tools have been developed for analyzing ribosome profiling data (for an overview see [2]), few address these challenges directly. Instead, tools typically rely on manual selection of read lengths and offsets or perform selection as part of an integrated pipeline with no data export options [3, 4, 5].

Here, we introduce Shoelaces, a software tool for processing ribosome profiling data. Shoelaces addresses the processing challenges by (1) utilizing a property of phasing, a strong 3-nucleotide periodicity of the reads stemming from coding regions [1, 6, 7] to filter genuine translating footprints and (2) calibrating P-site offsets based on metagene profiles over start or stop codons, stratified by footprint length [1, 8]. Shoelaces automatically selects these lengths and offsets, as well as offers batch-mode for processing multiple libraries in bulk.

The tool can be run in two modes: either using a graphical or command line interface. The graphical interface is accessible to users of all levels and guides the user through each processing step, allows for interactive adjustments and offers a range of extra visualization features on both gene/transcript or global level. The command line interface offers the same functionality as the graphical interface, without the interactivity, and can be easily integrated into automated processing pipelines.

Implementation

Shoelaces is implemented in Python3 and designed to run on Linux and MacOS operating systems. It relies on OpenGL for rendering graphics and PyQt5 for cross-platform graphical user interface. GUI is composed of a set of windows that user can easily rearrange by drag-and-drop to customize layout. The plots are interactive making the processing easily adjustable to specific needs. While primarily designed for the visualization features, Shoelaces can be also run in command line, making it easy to incorporate into processing pipelines. Shoelaces operates on common genomic formats (BAM, GTF, BED, wiggle), and stores settings in XML files, for maximum ease of use and reproducibility of analyses.



Leader: 535
Trailer: 39
Intron: 0
Non-coding: 0
CDS: 8000
Multimappers: 96
Sum: 8670
Strand: +

Figure 1: Shoelaces workflow. The toolkit accepts BAM and GTF files as input, filters reads, identifies translating lengths, determines P-site offsets and outputs tracks into wiggle format. Visual representation and summary statistics aid the processing steps.

Results and discussion

Data processing workflow

The workflow of Shoelaces is shown in Fig. 1. Shoelaces accepts standard genomic formats requiring alignment of ribosome profiling reads (BAM) and corresponding gene/ transcript annotations (GTF or BED). Shoelaces then guides the user through three main steps: (1) read filtering, (2) footprint identification and (3) P-site determination.

In the initial step Shoelaces filters reads from noise regions. Here, users can optionally include an additional annotation file with regions (such as e.g. rRNA or repetitive elements) which will be masked from all further analyses. Specific genes can also be deselected during this step if certain outliers are undesired.

In the following step, Shoelaces automatically determines the correct footprint lengths. This is based on the intrinsic 3-nucleotide periodicity characteristic of ribosome-derived fragments as opposed to reads originating from other processes [7]. The periodicity is detected using discrete Fourier transform (see below) over the CDS regions of annotated genes. Lengths displaying periodicity are selected for further analysis. The rest is classified as noise but is available for further analysis by the user.

Finally, for each footprint Shoelaces determines the codon that is actively translated. A length dependent P-site offset is calibrated using change point analysis (see below) over the distribution of footprints surrounding start and stop codons of annotated genes. Based on this, Shoelaces will automatically suggest offsets and provide plots of the summed footprints over start and stop codons of all genes. In addition, ribosome footprints are known to map preferentially to the first nucleotide in the codon [1] and Shoelaces therefore displays the fraction of reads falling into each reading frame. Manual adjustment is also possible if deemed necessary by the user.

After confirming the selection of the suggested footprint lengths and offsets, the user can export the ribosome coverage into flat file format (wiggle) for further downstream analysis. Optionally, different footprint lengths can be exported into separate files. Separation by length can be useful for more specific analysis, such as e.g. detection of conformational changes of ribosome at certain positions [9, 10].

To aid the researcher, the GUI produces summary statistics and counts for individual genes and transcripts, as well as for the whole library. It provides an overview over how many reads of a given length fall into different genomic regions (CDSs, 5' leaders, 3' trailers and introns) as well as how many footprints are found over non-coding transcripts or mapping to multiple positions in the genome. Users can update the statistics after read length and offset selection to see how they change. Together, these give an indication of the quality of the library and how well the reads represent genuine ribosome protected fragments.

Automatic selection of read lengths and offsets

An ideal-case scenario is presented in Fig. 2: the given footprint length is periodic (Fig. 2d), the metagene profiles have distinct peaks over start and before stop codons (Fig. 2a,b) and reads preferentially map to the first reading frame (Fig. 2c). However, library-specific biases can result in varying distributions of coding footprint lengths, as well as varying offsets (for various examples see Additional file 1). To take these biases into account, as well as to make processing large amounts of ribosome profiling data easy for the user, Shoelaces automatically suggests read lengths and offsets to be used.

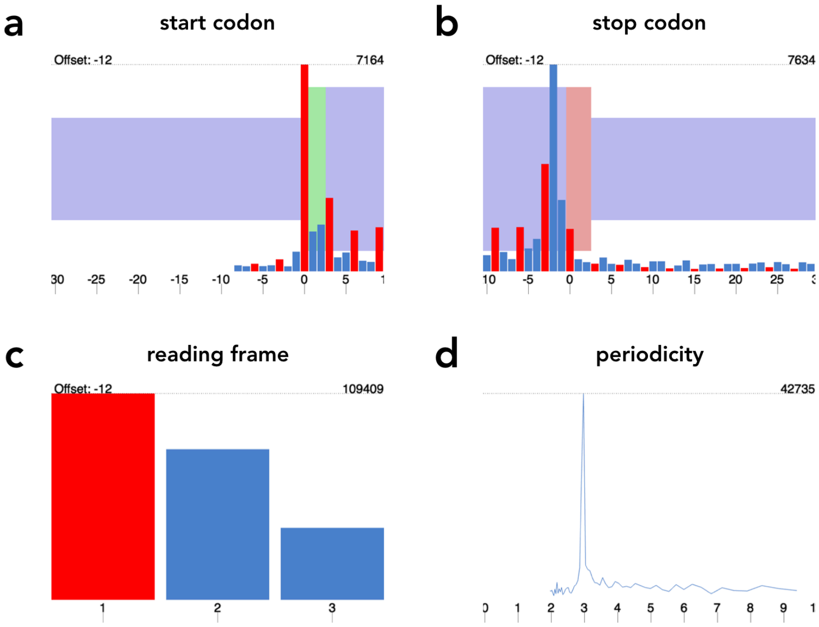


Figure 2: Read length and offset selection. In an ideal case scenario, the 3-nucleotide periodicity determines if the footprint length is coding (d), the peaks over start (a) and the last codon before stop (b) codons are used to calibrate offsets and most of the reads map to the first reading frame (c). Here, the plots demonstrate length 28 in human ribosome profiling sample (SRR493747, [15]). For more plots and datasets see Additional file 1.

Selection of periodic lengths

For each fragment length, the 5' ends of footprints mapping to the first 150 nucleotides of CDSs (by default from top 10% of protein-coding genes with highest coverage) are summed together. As the reads map preferentially to the first nucleotide of every codon, the periodic pattern will be conserved. The resulting vector is subject to discrete Fourier transform, and the fragment lengths whose highest amplitude corresponds to a period of 3 are considered to be periodic.

P-site determination

For each fragment length, the distribution of 5' ends of footprints surrounding start and stop codons (-30/+10 nucleotides) of protein-coding genes is calculated. The resulting window is subject to change point analysis, where for each adjacent position we calculate the difference of means. The maximum shift in means is assumed to correspond to the 5’ end of the footprints of initiating ribosomes and the distance from these to the P-site becomes the offset for that fragment length.

Visualization

Shoelaces also allows for visual inspection of coverage over individual genes (or group of genes) of interest. Users can manually zoom in/out to adjust the view, inspect the summary statistics with and without using offsets, and export high quality figures and tracks for further analysis and visualization.

Large-scale processing

For processing multiple libraries in bulk, a batch mode is available. For instance, for a number of same-batch libraries, one can be inspected visually, processing steps stored in an XML file and applied to the others. This additionally makes the processing easily reproducible later on. The processing can also be performed and fully automated from the command line allowing Shoelaces to be a part of a more comprehensive pipeline.

Analysis of human ribosome profiling data

We analyzed 79 libraries of human ribosome profiling data from 12 studies [11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22] and compared our read selection to the original, where applicable. Shoelaces retains up to 32% more data mapping to the coding regions of the genome (see Additional file 1, Table 1) than when originally processed. Non-periodic lengths are not selected, such as those that map primarily to 3' trailers, suggesting that they might originate from e.g. mRNA-binding proteins, abundant in 3' trailers, secondary structure or other sources of noise (Additional file 1, Figure 4).

Conclusions

Shoelaces aims for an intuitive and streamlined processing of libraries from different studies and treatments, making them comparable and analysis easily reproducible. The precision in bringing the data to sub-codon resolution is especially important in studies on translational efficiency of different codons, but also allows for detection of translational events such as ribosomal pausing [23], stop codon readthrough [3] or frameshifting [6]. The automation and batch processing facilitate dealing with large amounts of data, while visualization features add to user-friendliness and allow for more specific analyses. As we demonstrate on human ribosome profiling data, Shoelaces retains more reads mapping to coding regions than arbitrary manual processing. Overall, Shoelaces is a comprehensive tool for ribosome profiling data processing, and should prove useful to anyone interested in small or large-scale studies on ribosome profiling.

Availability of data and materials

The datasets analyzed in the current study are available in the Sequence Read Archive with accession numbers SRP038695 [11], SRP031501 [12], SRP002605 [13], SRP010679 [14], SRP012648 [15], SRP045257 [16], SRP014629 [17], SRP017263 [18], SRP053402 [19], SRP016143 [20], SRP029589 [21], SRP033369 [22]. The demo dataset is available together with the pipeline at https://bitbucket.org/valenlab/shoelaces.

Funding

This work was supported by the Bergen Research Foundation and the Norwegian Research Council (#250049).

Author's contributions

ÅB designed and implemented the software. KC implemented the algorithms for the method, tested the software, analyzed the data and wrote the manuscript. EV conceived the pipeline, guided the design and made critical revisions to the manuscript. All authors read and approved the final manuscript.

References

1. Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S., Weissman, J.S.: Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science 324(5924), 218–223 (2009)

2. Wang, H., Wang, Y., Xie, Z.: Computational resources for ribosome profiling: from database to web server and software. Brief Bioinform (2017)

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

4. Popa, A., Lebrigand, K., Paquet, A., Nottet, N., Robbe-Sermesant, K., Waldmann, R., Barbry, P.: Riboprofiling: a bioconductor package for standard ribo-seq pipeline processing. F1000Res 5, 1309 (2016)

5. Malone, B., Atanassov, I., Aeschimann, F., Li, X., Grosshans, H., Dieterich, C.: Bayesian prediction of rna translation from ribosome profiling. Nucleic Acids Res 45(6), 2960–2972 (2017)

6. Michel, A.M., Choudhury, K.R., Firth, A.E., Ingolia, N.T., Atkins, J.F., Baranov, P.V.: Observation of dually decoded regions of the human genome using ribosome profiling data. Genome Res 22(11), 2219–2229 (2012)

7. Bazzini, A.A., Johnstone, T.G., Christiano, R., Mackowiak, S.D., Obermayer, B., Fleming, E.S., Vejnar, C.E., Lee, M.T., Rajewsky, N., Walther, T.C., Giraldez, A.J.: Identification of small orfs in vertebrates using ribosome footprinting and evolutionary conservation. EMBO J 33(9), 981–993 (2014)

8. Woolstenhulme, C.J., Guydosh, N.R., Green, R., Buskirk, A.R.: High-precision analysis of translational pausing by ribosome profiling in bacteria lacking efp. Cell Rep 11(1), 13–21 (2015)

9. Giess, A., Jonckheere, V., Ndah, E., Chyzynska, K., Van Damme, P., Valen, E.: Ribosome signatures aid bacterial translation initiation site identification. BMC Biol 15(1), 76 (2017)

10. Lareau, L.F., Hite, D.H., Hogan, G.J., Brown, P.O.: Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mrna fragments. Elife 3, 01257 (2014)

11. Andreev, D.E., O'Connor, P.B.F., Fahey, C., Kenny, E.M., Terenin, I.M., Dmitriev, S.E., Cormican, P., Morris, D.W., Shatsky, I.N., Baranov, P.V.: Translation of 5' leaders is pervasive in genes resistant to eif2 repression. Elife 4, 03971 (2015)

12. Gonzalez, C., Sims, J.S., Hornstein, N., Mela, A., Garcia, F., Lei, L., Gass, D.A., Amendolara, B., Bruce, J.N., Canoll, P., Sims, P.A.: Ribosome profiling reveals a cell-type-specific translational landscape in brain tumors. J Neurosci 34(33), 10924–10936 (2014)

13. Guo, H., Ingolia, N.T., Weissman, J.S., Bartel, D.P.: Mammalian micrornas predominantly act to decrease target mrna levels. Nature 466(7308), 835–840 (2010)

14. Hsieh, A.C., Liu, Y., Edlind, M.P., Ingolia, N.T., Janes, M.R., Sher, A., Shi, E.Y., Stumpf, C.R., Christensen, C., Bonham, M.J., Wang, S., Ren, P., Martin, M., Jessen, K., Feldman, M.E., Weissman, J.S., Shokat, K.M., Rommel, C., Ruggero, D.: The translational landscape of mtor signalling steers cancer initiation and metastasis. Nature 485(7396), 55–61 (2012)

15. Ingolia, N.T., Brar, G.A., Rouskin, S., McGeachy, A.M., Weissman, J.S.: The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mrna fragments. Nat Protoc 7(8), 1534–1550 (2012)

16. Ingolia, N.T., Brar, G.A., Stern-Ginossar, N., Harris, M.S., Talhouarne, G.J.S., Jackson, S.E., Wills, M.R., Weissman, J.S.: Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. Cell Rep 8(5), 1365–1379 (2014)

17. Lee, S., Liu, B., Lee, S., Huang, S.-X., Shen, B., Qian, S.-B.: Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. Proc Natl Acad Sci U S A 109(37), 2424–32 (2012)

18. Liu, B., Han, Y., Qian, S.-B.: Cotranslational response to proteotoxic stress by elongation pausing of ribosomes. Mol Cell 49(3), 453–463 (2013)

19. Sidrauski, C., McGeachy, A.M., Ingolia, N.T., Walter, P.: The small molecule isrib reverses the effects of eif2alpha phosphorylation on translation and stress granule assembly. Elife 4 (2015)

20. 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., Weissman, J.S.: Decoding human cytomegalovirus. Science 338(6110), 1088–1093 (2012)

21. Stumpf, C.R., Moreno, M.V., Olshen, A.B., Taylor, B.S., Ruggero, D.: The translational landscape of the mammalian cell cycle. Mol Cell 52(4), 574–582 (2013)

22. Subtelny, A.O., Eichhorn, S.W., Chen, G.R., Sive, H., Bartel, D.P.: Poly(a)-tail profiling reveals an embryonic switch in translational control. Nature 508(7494), 66–71 (2014)

23. Li, G.-W., Oh, E., Weissman, J.S.: The anti-shine-dalgarno sequence drives translational pausing and codon choice in bacteria. Nature 484(7395), 538–541 (2012)

Supplementary material:

Additional file 1 — Analysis examples
Figure 1-3: Three different examples of offset selection (PDF file) for human ribosome profiling datasets: SRR493747 [15], treated with harringtonine and cyclohexamide; SRR1039861 [22], treated with cyclohexamide; SRR592961 [20], no drug. Table 1: Comparison of selected footprint lengths as originally in human ribosome profiling studies and Shoelaces. Figure 4: Comparison of reads mapping to different parts of transcript as selected by Shoelaces and the original manual selection (SRR493747 [15]).