Bioinformatics, 31(15), 2015, 2586–2588
doi: 10.1093/bioinformatics/btv174
Advance Access Publication Date: 25 March 2015
Applications Note

OXFORD

Gene expression

RiboTools: a Galaxy toolbox for qualitative ribosome profiling analysis

Rachel Legendre, Agnès Baudin-Baillieu, Isabelle Hatin and Olivier Namy*

Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Université Paris-Sud, Bat 400, 91400 Orsay, France

*To whom correspondence should be addressed.

Associate Editor: Ziv Bar-Joseph

Received on October 20, 2014; revised on January 23, 2015; accepted on March 22, 2015

Abstract

Motivation: Ribosome profiling provides genome-wide information about translational regulation. However, there is currently no standard tool for the qualitative analysis of Ribo-seq data. We present here RiboTools, a Galaxy toolbox for the analysis of ribosome profiling (Ribo-seq) data. It can be used to detect translational ambiguities, stop codon readthrough events and codon occupancy. It provides a large number of plots for the visualisation of these events.

Availability and implementation: RiboTools is available from https://testtoolshed.g2.bx.psu.edu/view/rlegendre/ribo_tools as part of the Galaxy Project, under the GPLv3 licence. It is written in python2.7 and uses standard python libraries, such as matplotlib and numpy.

Contact: olivier.namy@igmors.u-psud.fr

Supplementary Information: Supplementary data are available from *Bioinformatics* online.

1 Introduction

Translation fidelity is essential for the correct decoding of genetic information. However, translation fidelity is not maximal, and several genes use alternative mechanisms to regulate their expression (Eswarappa *et al.*, 2014; Namy *et al.*, 2004). Such recoding events are difficult to identify by bioinformatics methods and their incorrect annotation in databases is still frequent. Genome-wide translational analyses, such as ribosome profiling, can identify many features affecting translational quality, such as ribosomal pauses, dual coding regions, translation start sites at AUG or non-AUG codons or unconventional translational decoding events, the translation of non-coding RNA and translational ambiguities (Baudin-Baillieu *et al.*, 2014; Ingolia *et al.*, 2014; Ingolia *et al.*, 2011; Menschaert *et al.*, 2013; Michel *et al.*, 2012; Wan and Qian, 2014).

Ribosome profiling has been widely adopted by the scientific community, because it has revolutionised the ways in which translation can be studied (Ingolia *et al.*, 2014). It can be used for both quantitative and qualitative analyses of translation. Quantitative analysis is carried out with standard RNA-seq scripts and is easy to implement (Olshen *et al.*, 2013). An online genome browser is also available for the visualisation of many ribosome profiling data (Michel *et al.*, 2014). However, qualitative analysis is much more challenging, because there is no standard analysis technique available. Correct analysis requires the development of specific scripts, as most publications describe only briefly the way in which the data should be analysed. A better standardization of ribosome profiling approaches is urgently needed to limit misinterpretations (Artieri and Fraser, 2014; Gerashchenko and Gladyshev, 2014).

We describe here RiboTools, an efficient Galaxy package that can be used for primary analyses of Ribo-seq data, and to address common issues, such as the identification of translational ambiguities, or of stop codon readthrough events, and the codon occupancy of ribosomal A, P, and E-sites. RiboTools provides user-friendly publication-grade graphical results (html report). Scripts are available from the Galaxy bioinformatics platform (Goecks *et al.*, 2010) via ToolShed. RiboTools will be useful to the research community as it facilitates the complete qualitative analysis of ribosome profiling data.

2 Programs

We assume that ribosome-protected fragments (footprints) have already been mapped to a reference genome and that the data are

© The Author 2015. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

available in a binary version of sequence map/alignment (BAM) format. To this end, we provide a supplementary file containing a tutorial explaining how to obtain the BAM file from the FastaQC file. The BAM file must be sorted, but no subalignment extraction is required before its use. All the scripts use two automatic alignment filters: uniquely mapped footprints only, and footprints mapped with high quality (MAPQ > 20). GFF3 files are required for feature annotation.

2.1 Primary analysis and k-mer choice

Kmer can be used to choose the best length of footprint, k-mer, with the largest number of in-frame footprints. Two kinds of analysis are performed: (i) size distribution of the mapped footprints present in the BAM file; (ii) classification of the category of each footprint, within the three possible reading frames from the translational start and stop sites annotated in the GFF3 file (Supplementary Figure S1A).

2.2 Detection of translational ambiguities

Translational ambiguities are coding sequences for which footprints are simultaneously present in several frames. *Frame* uses the k-mer selected as described above to detect such rare events, by analysing the out-of-frame footprints in each annotated coding sequence (CDS) presents in the GFF3 file.

For this analysis, we divide CDS into segments of a size defined by the user. We then compare k-mer distribution in each segment with the global footprint distribution in the three possible reading frames (obtained by *kmer*, Supplementary Figure S1A) and we report genes with unconventional distributions.

This tool also generates an html report and a subprofile plot of genes with translational ambiguities (Supplementary Figure S1B).

2.3 Identification of readthrough events

Stop codon readthrough corresponds to ribosomes translating part of the 3′UTR. Such events have been reported to occur in complex genomes (Eswarappa *et al.*, 2014; Loughran *et al.*, 2014; Namy *et al.*, 2003; Schueren *et al.*, 2014). We use the parameters described by Dunn and collaborators (Dunn *et al.*, 2013) to identify potential stop codon readthrough events. *Stop_supp* considers all the footprints present in the BAM file. Potential readthrough is considered to occur if: (i) there are footprints after the stop codon of the CDS; (ii) there are footprints overlapping the next in-frame stop codon; (iii) there is no methionine codon in the next five codons downstream from the stop codon of the CDS; (iv) there is homogeneous coverage of the extension.

Stop codon readthrough is estimated by calculating the ratio of the number of footprints in the C-terminal extension to the number in the CDS. Ribosome density footprints are estimated in RPKM (reads per kilobase per million). We control for variability due to stop codon peaks, by excluding footprints mapping to stop codons from the calculation of RPKM.

2.4 Codon occupancy at a specific ribosome site

The *codon_density* tool determines whether codon occupancy differs between two sets of conditions, by calculating the numbers of each codon in the theoretical position of the ribosome site selected by the user. Classically, the A, P and E-sites are located 15, 12 and 9 nucleotides downstream from the start of the footprint. However, this parameter can be modified by the user. This tool uses the footprint size defined by the user with *Kmer*. *Codon_density* plots codon occupancy for each set of conditions and evaluates differential occupancy with a chi-squared test. Codons are also grouped by amino acid in another plot, for visualisation of the possible global effects (Supplementary Figure S1C).

3 Example of application

We recently used RiboTools to analyse the translational impact of the [PSI+] prion in yeast (Baudin-Baillieu *et al.*, 2014). Ribosomal footprints of different lengths may be obtained, due to the variability of digestion efficiency. *Kmer* can be used to visualise this diversity and to identify the best in-frame subfamily. For example, in yeast, more than 80% of the 28-mers are in-frame (Supplementary Figure S1A).

Using *Frame*, we were able to identify 119 genes with translational ambiguities, revealing, for the first time, an unexpected effect of [PSI+]. Moreover, *Stop_supp* identified 205 potential readthrough events generated by [PSI+].

Codon_density revealed no significant differences in the codon occupancy of ribosomal sites other than for stop codons, consistent with a larger number of ribosomes translating the stop codon and continuing downstream (Supplementary Figure S1C). We checked that RiboTools was suitable for use with other datasets, by also analysing published ribosome profiling data from other groups working on yeast, *Drosophila* and mouse. We obtained results similar to those published (see Supplementary Figures S2A and S2B).

4 Conclusion

The methods provided by RiboTools are designed for the accurate analysis of k-mer length distribution, translation ambiguities and translation readthrough events. They evaluate codon occupancy at a specific ribosome site on the basis of Ribo-seq data. These tools have been tested and the results support their statistical and computational properties. We report here an integrated open-source Galaxy tool providing the broader research community with access to these methods. RiboTools is available from the public ToolShed repository.

Acknowledgements

We thank Alban Ott, Vladimir Daric, Claire Toffano-Nioche and Coline Billerey for helpful discussions. We also thank the eBio platform for bioinformatics support.

Funding

This work and R.L are supported by a grant from the ANR (ANR-2011-BSV6-011 and ANR-13-BSV8-0012-02) to O.N.

Conflict of Interest: none declared.

References

Artieri,C.G. and Fraser,H.B. (2014) Accounting for biases in riboprofiling data indicates a major role for proline in stalling translation. *Genome Res.*, **24**, 2011–2021.

Baudin-Baillieu,A. *et al.* (2014) Genome-wide translational changes induced by the prion [PSI+]. *Cell Rep.*, **8**, 439–448.

Dunn,J.G. *et al.* (2013) Ribosome profiling reveals pervasive and regulated stop codon readthrough in *Drosophila melanogaster*. *eLife* **2**, e01179.

Eswarappa,S.M. *et al.* (2014) Programmed translational readthrough generates antiangiogenic VEGF-Ax. *Cell* **157**, 1605–1618.

Gerashchenko,M.V. and Gladyshev,V.N. (2014) Translation inhibitors cause abnormalities in ribosome profiling experiments. *Nucleic Acids Res.*, **42**, e134.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Goecks,J. *et al*. (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, **11**, R86.

Ingolia,N.T. *et al*. (2014) Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep.*, **8**, 1365–1379.

Ingolia,N.T. *et al*. (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*, **147**, 789–802.

Loughran,G. *et al*. (2014) Evidence of efficient stop codon readthrough in four mammalian genes. *Nucleic Acids Res.*, **42**, 8928–8938.

Menschaert,G. *et al*. (2013) Deep proteome coverage based on ribosome profiling aids mass spectrometry-based protein and peptide discovery and provides evidence of alternative translation products and near-cognate translation initiation events. *Mol. Cell. Proteomics: MCP*, **12**, 1780–1790.

Michel,A.M. *et al*. (2012) Observation of dually decoded regions of the human genome using ribosome profiling data. *Genome Res.*, **22**, 2219–2229.

Michel,A.M. *et al*. (2014) GWIPS-viz: development of a ribo-seq genome browser. *Nucleic Acids Res.*, **42**, D859–D864.

Namy,O. *et al*. (2003) Identification of stop codon readthrough genes in *Saccharomyces cerevisiae*. *Nucleic Acids Res.*, **31**, 2289–2296.

Namy,O. *et al*. (2004) Reprogrammed genetic decoding in cellular gene expression. *Molecular Cell.*, **13**, 157–168.

Olshen,A.B. *et al*. (2013) Assessing gene-level translational control from ribosome profiling. *Bioinformatics*, **29**, 2995–3002.

Schueren,F. *et al*. (2014) Peroxisomal lactate dehydrogenase is generated by translational readthrough in mammals. *Elife*, **23**, e03640.

Wan,J. and Qian,S.B. (2014) TISdb: a database for alternative translation initiation in mammalian cells. *Nucleic Acids Res.*, **42**, D845–D850.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016