SOFTWARE TOOL ARTICLE

RiboProfiling: a Bioconductor package for standard Ribo-seq pipeline processing [version 1; referees: 3 approved]

Alexandra Popa, Kevin Lebrigand, Agnes Paquet, Nicolas Nottet, Karine Robbe-Sermesant, Rainer Waldmann, Pascal Barbry

Institut de Pharmacologie Mol´eculaire et Cellulaire, University Nice Sophia Antipolis and CNRS, Sophia- Antipolis, 06560, France

v1 **First published:** 09 Jun 2016, **5**:1309 (doi: 10.12688/f1000research.8964.1)
Latest published: 09 Jun 2016, **5**:1309 (doi: 10.12688/f1000research.8964.1)

Abstract

The ribosome profiling technique (Ribo-seq) allows the selective sequencing of translated RNA regions. Recently, the analysis of genomic sequences associated to Ribo-seq reads has been widely employed to assess their coding potential. These analyses led to the identification of differentially translated transcripts under different experimental conditions, and/or ribosome pausing on codon motifs. In the context of the ever-growing need for tools analyzing Ribo-seq reads, we have developed 'RiboProfiling', a new Bioconductor open-source package. 'RiboProfiling' provides a full pipeline to cover all key steps for the analysis of ribosome footprints. This pipeline has been implemented in a single R workflow. The package takes an alignment (BAM) file as input and performs ribosome footprint quantification at a transcript level. It also identifies footprint accumulation on particular amino acids or multi amino-acids motifs. Report summary graphs and data quantification are generated automatically. The package facilitates quality assessment and quantification of Ribo-seq experiments. Its implementation in Bioconductor enables the modeling and statistical analysis of its output through the vast choice of packages available in R. This article illustrates how to identify codon-motifs accumulating ribosome footprints, based on data from *Escherichia coli*.

This article is included in the Bioconductor channel.

This article is included in the RPackage channel.

Open Peer Review

Referee Status: ✓ ✓ ✓

	Invited Referees 1	2	3
version 1 published 09 Jun 2016	✓ report	✓ report	✓ report

1 **Olivier Namy**, University of Paris-Sud France, **Pierre Bertin**, Université Paris-Sud France

2 **Audrey M Michel**, University College Cork Ireland

3 **Ghislain Bidaut**, Inserm, Aix-Marseille Université, CNRS and Institut Paoli-Calmettes, Centre de Recherche en Cancérologie de Marseille France

Discuss this article

Comments (0)

Corresponding author: Pascal Barbry (barbry@ipmc.cnrs.fr)

How to cite this article: Popa A, Lebrigand K, Paquet A *et al.* **RiboProfiling: a Bioconductor package for standard Ribo-seq pipeline processing [version 1; referees: 3 approved]** *F1000Research* 2016, **5**:1309 (doi: 10.12688/f1000research.8964.1)

Grant information: This work was developed by the Functional Genomics Platform at Nice Sophia Antipolis, a partner of the National Infrastructure France Génomique (ANR-10-INBS-09-03 and ANR-10-INBS-09-02) and PB's group, thanks to supports by the Cancéropôle PACA and Commissariat aux Grands Investissements. RW was supported by Fondation ARC pour la recherche sur le cancer (SFI20121205973), and PB by ANR (ANR-12-BSVI-0023-02), Fondation pour la Recherche Médicale (DEQ20130326464) and labex Signalife (ANR-11-LABX-0028-01).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: No competing interests were disclosed.

First published: 09 Jun 2016, **5**:1309 (doi: 10.12688/f1000research.8964.1)

Introduction

Ribosome profiling (Ribo-seq) is a recently developed high throughput sequencing technique (Ingolia *et al.*, 2009) that allows the identification of RNA fragments resistant to RNAse digestion. Fragments mainly correspond to coding sequences protected against RNAse digestion by translating ribosomes. Ribo-seq data have been widely used to assess the translational status of open reading frames (ORFs) (Bazzini *et al.*, 2014; Fields *et al.*, 2015; Ingolia *et al.*, 2009; Popa *et al.*, 2016), and to identify ORFs differentially translated between experimental conditions (Schafer *et al.*, 2015).

Ribo-seq bioinformatics analyses comprise the selection of reads consistent with ribosome footprints, a recalibration of the read start or end (5' or 3' extremity of the read depending on the RNA digestion step) to the peptidyl site (P-site) position of the ribosome, and quantification of the reads on specific features of interest (i.e. transcript, codons, multi-codons motifs). Several tools for processing ribosome profiling data have previously been proposed. RiboTools' (Legendre *et al.*, 2015) and 'RUST' (O'Connor *et al.*, 2015)) were developed in python, 'riboSeqR' (Hardcastle, 2014) corresponds to an R package. Each of the above software integrates some, but not all, of the functions necessary for a standard Ribo-seq workflow from reads to quality assessment, recalibration and quantification. A previous effort to group the different approaches for Ribo-seq analyses has been developed with the Galaxy instance RiboGalaxy (Michel *et al.*, 2016).

These tools have been developed to answer specific questions related to ribosome occupancy: normalization of Ribo-seq reads ('RUST'), detection and characterization of reading frame usage ('riboSeqR'), or irregular translational behavior such as translational ambiguities (Ribo-seq footprints in different phases) and stop-codon read through ('RiboTools'). However, prior to any advanced Ribo-seq data processing, it is necessary to have standard pipelines for quality assessment of the experiments and specific ribosome footprint assignment to sequences.

We implemented the features for a standard Ribo-seq workflow in a pipeline entitled 'RiboProfiling'. The pipeline takes an alignment file (BAM) as input, performs identification of the read offset, generates transcript and (multi-) codon coverage quantification data, and performs statistical analyses as well as graphical representations. Our pipeline is, to our knowledge, the most complete integration of a ribosome profiling standard analysis pipeline in an unique R framework. This includes the crucial step prior to quantifying ribosome footprints that consists in identifying the offset between Ribo-seq reads and the P-site of the ribosome. We have given special attention to this step as it is essential to correctly associate ribosome footprints with codon resolution. Depending on the RNA digestion protocol, the assignment of the ribosome must be made specifically either to the 5', 3' or the center of the read. To our knowledge, there is only one implementation for the determination of ribosome offset that was published so far. It was proposed in 'riboSeqR' as a metagene plot and the determination of the offset was only possible from the 5' read end and for a particular read length. Our package offers several options, to compute the offset and recalibrate reads based either on the 5' or the 3' read ends. RiboProfiling also enables the graphical representation of ribosome density around the Translation Start Site (TSS) for multiple read lengths. This option allows to perform analysis for a single read length or on the merge of several read lengths. It enables to group all lengths sharing a same and unique offset value. The R/Bioconductor implementation provides an easy-to-use comprehensive set of functions that requires a minimal knowledge in R programming. The package contains a function entitled 'riboSeqFromBAM' that treats multiple Ribo-seq BAM files in parallel. The automated workflow generates report summary graphs and data quantification.

We illustrate the main features of the package using a Ribo-seq control sample from murine ES cells (GSM1655059), taken from a recently published ribosome profiling study using translation inhibitors (Popa *et al.*, 2016). We then detail the analysis of ribosome accumulation on certain codons and tri-peptide motifs on a public dataset in *Escherichia coli* (GSE64488) (Woolstenhulme *et al.*, 2015). The script for performing all these analyses is publicly available.

Methods

Requirements

The 'RiboProfiling' package, v.1.2.0 can be used with R 3.3.0 and Bioconductor version 3.3. The script needs a minimum memory limit in R of 3 Gb when analyzing tripeptide motifs. The package starts from alignment BAM files, from either Ribo-seq or RNA-seq experiments. We have validated BAM files from bowtie/tophat, Hisat2, STAR, and Lifescope (Solid), both single- or paired-end (for RNA-seq reads). Reads from rRNA, tRNA, and PCR duplicates (if unique molecular identifiers are available) should be removed from the BAM files before starting the analysis (see package vignette for details).

RiboProfiling package

All analyses can be performed either through a call to a function called 'riboSeqFromBAM', or through a step by step approach. Figure 1 describes the workflow of the package starting from BAM files with reads mapped to the genome of interest. The first step in processing Ribo-seq reads is to select only those with match lengths compatible with standard ribosome footprints. The function 'histMatchLength' allows the visual inspection of read match sizes distribution across the BAM file (Figure 2.A), which should be enriched in reads of sizes between 20 to 40 nucleotides (Ingolia *et al.*, 2009; Popa *et al.*, 2016).

A second particularity in the handling of Ribo-seq data comes from the shift existing between the extremities of the read (i.e. 5' or 3') and the P-site position of the ribosome. Our package allows the identification of an offset from the 5' end of the read, but also from the 3' end. The function 'readStartCov' computes the read frequency distribution centered on the translation start site (TSS) of the most expressed protein coding transcripts (by default the 3% most expressed). Based on this frequency distribution, the 'plotSummarizedCov' function enables the visual quantification of the offset between the reads and the ribosome P-site (Figure 2.B).

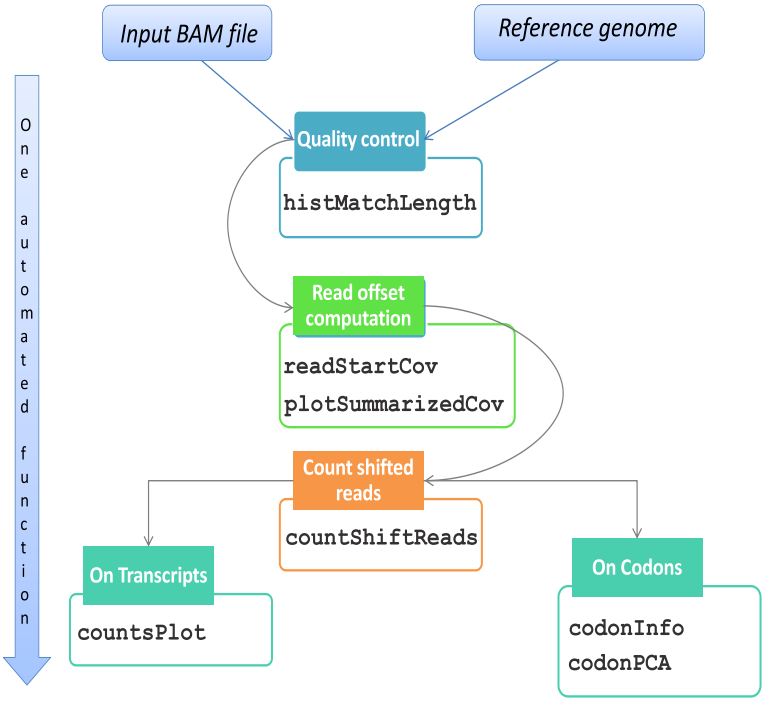

Figure 1. Workflow of 'RiboProfiling' Ribo-seq analysis from BAM to quantification on genomic features and codon motifs.

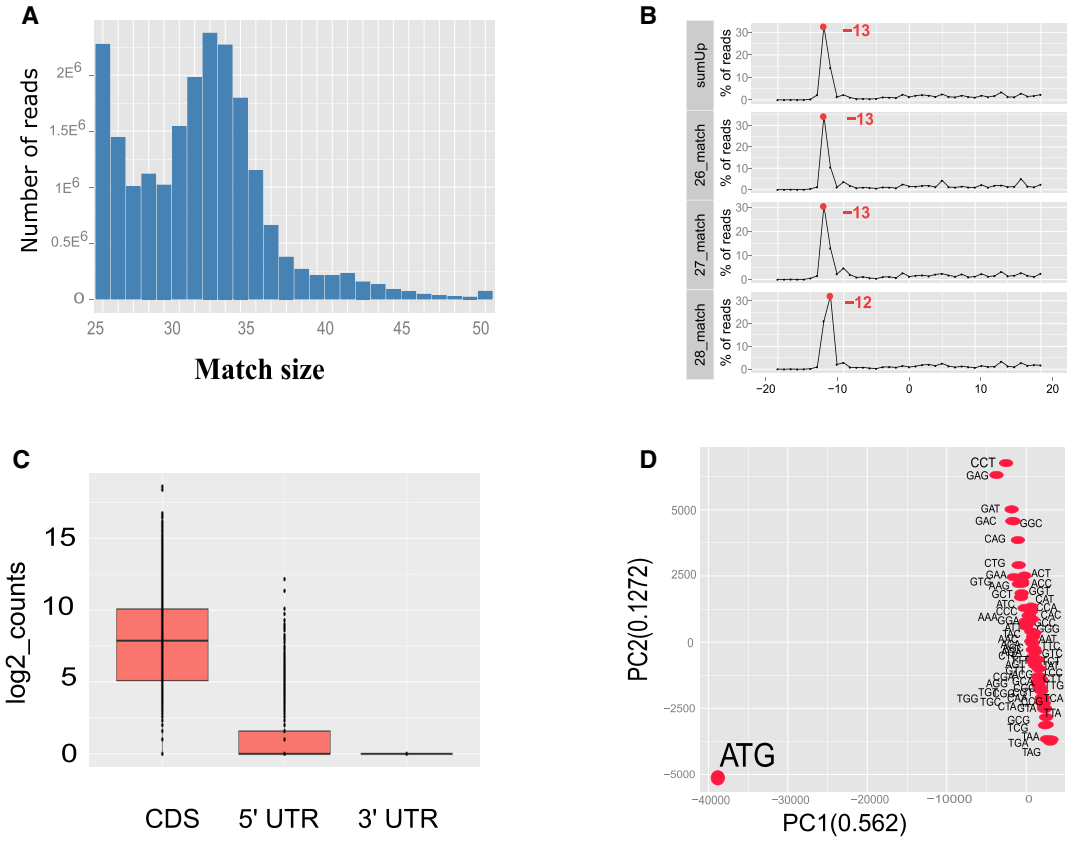

Figure 2. Graphical output of 'RiboProfiling' package on the GSM1655059 dataset. A, Histogram of read length distribution with 'histMatchLength' function. **B**, Frequency of reads around the TSS for reads sizes 26 to 28; this graph points to an offset of 13 bp between the read 5' end and the ribosome P-site. **C**, Boxplots of Ribo-seq read coverage on the CDS, 5' UTR and 3' UTR regions of protein coding genes. **D**. PCA analysis of Ribo-seq coverage on codons.

In our Ribo-seq example, the 5' read end is shifted 13 bp from the TSS. The innovation of this feature consists in the visualization of read lengths independently and as a summary figure.

When computed, the offset can be applied on all reads based on the transcript referential with the function 'countShiftReads' and coverage on three different sequence features: 5'-UTR, coding sequences (CDS), and 3'-UTR (Figure 3.A). In Figure 2.C we observe that the majority of reads accumulate on the CDS of protein coding sequences and are practically lacking in the 3' non-coding UTR regions. Ribosome footprints are also detectable in the 5' UTR regions of protein coding genes, suggesting either the presence of coding upstream ORFs (Popa *et al.*, 2016) or possible confounding information from missing annotations.

Finally, our package provides quantification of ribosome footprints at codon resolution (Figure 3.B). A PCA analysis of codon occupancy can be performed and several graphical functions are implemented. In Figure 2.D we employed the 'codonInfo' and 'codonPCA' functions to analyze the codons accumulating ribosome footprints. As expected, the codon ATG is the most discriminant codon in the PCA analyses, since ribosome accumulation peaks are observed at the start codon of coding regions. Detailed descriptions with examples of the pipeline from BAM files to Ribo-seq reads quantification and processing are available in the vignette of our package: https://www.bioconductor.org/packages/release/bioc/vignettes/RiboProfiling/inst/doc/RiboProfiling.pdf.

Analysis of ribosome stalling on sequence motifs

'RiboProfiling' can also be useful for analyzing ribosome occupancy on multi-codons motifs. Codons accumulating ribosome footprints are indicative of slowed ribosome progression (stalling) during the translation elongation process. The 'RiboProfiling' package offers several features for quantifying footprint accumulation on sequence motifs (ranging from one to three consecutive codons), performs principal component analyses, and allows graphical representation of those data.

To illustrate how 'RiboProfiling' can be used to explore the influence of sequence motifs (in this case tri-amino-acid sequences) on ribosome pausing, we analyzed an *Escherichia coli* Ribo-seq dataset (Woolstenhulme *et al.*, 2015). We downloaded, filtered and mapped the reads of an efp-knockout sample (ΔEfp2, GSE64488), the elongation factor EFP being essential for the translation of polyproline motifs. Uniquely mapped reads from the resulting BAM file were analyzed with our 'RiboProfiling' package. After quality assessment of the reads size distribution, we quantified the offset between the 3' end of the reads and the TSS for different alignment match sizes (Figure 4.A). We can clearly observe the 15 nucleotides offset that was reported by the authors for reads with alignment sizes >= 30 nucleotides. Smaller match lengths exhibited either a strong variation in the distribution of reads around the TSS (i.e. the 29 mers), or a different offset (i.e. offset of 20 for 21mers) (Figure 4.A). We selected the reads with alignment match sizes between 30 and 40 nucleotides and quantified codon coverage by positioning the ribosome P-site 15 nucleotides upstream of the read 3' extremity.

An important stalling has been reported after the incorporation of two consecutive prolines (Pro-Pro) in the peptide chain. This stalling is highly dependent on the nature of the codon that follows the aminoacyl-tRNA reacting at the A-site (Woolstenhulme *et al.*, 2015). Following the article's analysis of stalling at PP(X) (Proline – Proline – 3rd codon) motifs, we used the 'countShiftReads' and 'codonInfo' functions to quantify the ribosome footprints on these motifs, in the ΔEfp2 sample. We then

A

	gene	chr	strand	transc_genomic_start	transc_genomic_end	transc_length	orf_start	orf_end
uc007afn.1	chr1	+	5083173	5162549	1976	141	1592
uc007aia.2	chr1	+	12692430	12860372	4625	483	3095
uc007aib.2	chr1	+	12692430	12860372	4530	388	3000
uc007aic.1	chr1	+	12718545	12822916	2660	467	1897
uc007aid.1	chr1	+	12718545	12860372	4609	467	3079

	orf_length	CDS_counts	fiveUTR_counts	threeUTR_counts
	1452	1804	88	2
	2613	2810	859	43
	2613	2810	793	43
	1431	1776	472	0
	2613	2810	472	43

B

	AAA	AAC	AAG	AAT	ACA	ACC	...	TGG	TGT	TTA	TTC	TTG	TTT
uc007afn.1	21	8	11	12	6	2		9	7	4	6	10	11
uc007aia.2	25	33	40	21	17	13	...	13	13	8	19	7	15
uc007aib.2	25	33	40	21	17	13		13	13	8	19	7	15
uc007aic.1	12	18	17	11	9	9		6	4	3	9	3	10
uc007aid.1	25	33	40	21	17	13		13	13	8	19	7	15

Figure 3. Output tables of 'RiboProfiling' package on the GSM1655059 dataset. A, Quantification of Ribo-seq read coverage on the CDS, 5' UTR and 3' UTR regions of protein coding genes. **B**. Quantification of Ribo-seq coverage on codons.

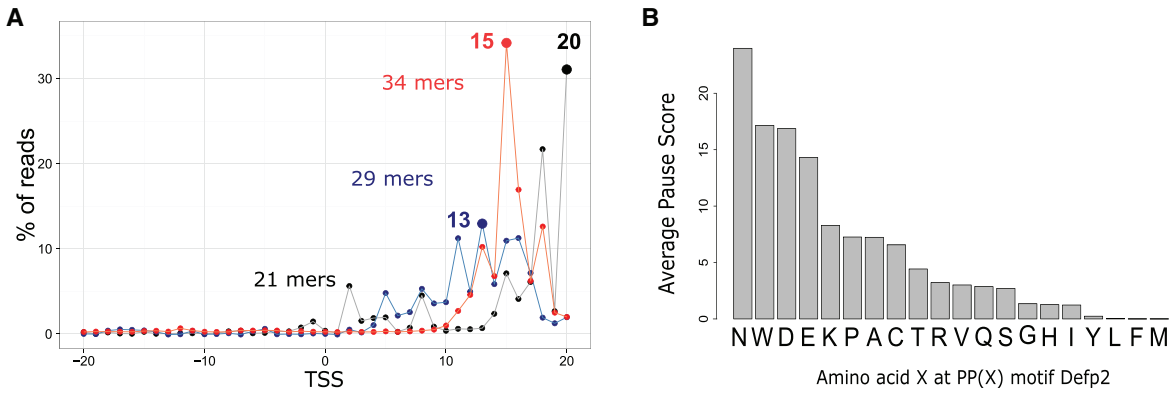

Figure 4. A, Sample ΔEfp2: percentage of read 3' end coverage 20 nucleotides around the TSS. Three match sizes are represented: 21, 29 and 34 mers. **B**, Barplot of the average ribosome occupancy on PP(X) motifs in the ΔEfp2 sample, where X is any of the 20 possible amino acids.

computed the pause score for all 20 possible PP(X) combinations in each ORF with more than 20 ribosome footprints, independently:

$$PauseScore_{PPX|ORF} = \frac{\frac{Reads_{PPX|ORF}}{Reads_{ORF}}}{\frac{Nbr_{PPX|ORF}}{Length_{ORF}}}$$

where, $Reads_{PPX|ORF}$ is the ribosome density of motif PPX in a given ORF; $Reads_{ORF}$ is the ribosome density on the ORF; $Nbr_{PPX|ORF}$ is the number of time a given PPX motif is present in the ORF; $Length_{ORF}$ is the total length of the ORF. We averaged the ribosome occupancy for each PP(X) motif on all ORFs. Figure 4.B shows a strong stalling in sample ΔEfp2 when the ribosome encounters PPN, PPW, PPD, in agreement with their previous identification as pause sites *in Escherichia coli* (Woolstenhulme *et al.*, 2015). A step by step R script implementing this entire analysis is provided at:

http://genomique.info/data/public/RiboProfiling/scriptWoolstenhulme_Defp2.R.

Summary

Our 'RiboProfiling' Bioconductor package offers a collection of tools for Ribo-seq data analysis. It provides an unique, straightforward R implementation of a ribosome profiling pipeline from BAM, to P-site calibration, quantification of reads on sequence features, and codon coverage. The packages' graphical features offer quality assessment and result representation across the analyses. Following the overview of Ribo-seq experiments with 'RiboProfiling, the output tables can then be easily integrated into more specialized dowstream analyses, either using more specialisez riboseq tools such as (XXX, YYY) or directly within R.

We highlighted here the features of our package in characterizing ribosome stalling at sequence motifs along ORFs based on an example dataset from Woolstenhulme *et al.* (Woolstenhulme *et al.*, 2015). The workflow we propose for the analysis of ribosome occupancy on codon motifs using the 'RiboProfiling' package will most surely prove an useful asset in the context of recent ribosome profiling applications such as the detection of tumor sensitivity to differential amino acid depletion (Loayza-Puch *et al.*, 2016).

Software availability

The package is built in R (>=3.3.0) and freely available from Bioconductor website

1. https://www.bioconductor.org/packages/release/bioc/html/RiboProfiling.html

2. We provide an associated-script to the analyses in this paper at http://genomique.info/data/public/RiboProfiling/scriptWoolstenhulme_Defp2.R

   Zenodo: scriptWoolstenhulme_Defp2.R, doi: 10.5281/zenodo.54567, (Popa A, 2016)

3. GPL-3 licence

Author contributions

R.Waldmann, A.Popa, A.Paquet, and P. Barbry discussed the functions implemented in the package. A.Popa and K.Lebrigand developed the package. N.Nottet downloaded and performed the secondary analysis of Ribo-seq public datasets. A.Popa, A. Paquet, P. Barbry, R. Waldmann, K. Robbe-Sermesant wrote the article.

Competing interests

No competing interests were disclosed.

Grant information

This work was developed by the Functional Genomics Platform at Nice Sophia Antipolis, a partner of the National Infrastructure

France Génomique (ANR-10-INBS-09-03 and ANR-10-INBS-09-02) and PB's group, thanks to supports by the Cancéropôle PACA and Commissariat aux Grands Investissements. RW was supported by Fondation ARC pour la recherche sur le cancer (SFI20121205973), and PB by ANR (ANR-12-BSVI-0023-02), Fondation pour la Recherche Médicale (DEQ20130326464) and labex Signalife (ANR-11-LABX-0028-01).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Bazzini AA, Johnstone TG, Christiano R, *et al.*: **Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation.** *EMBO J.* 2014; **33**(9): 981–993.
PubMed Abstract | Publisher Full Text | Free Full Text

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

Hardcastle TJ: **riboSeqR: Analysis of sequencing data from ribosome profiling experiments.** 2014.
Reference Source

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

Legendre R, Baudin-Baillieu A, Hatin I, *et al.*: **RiboTools: a Galaxy toolbox for qualitative ribosome profiling analysis.** *Bioinformatics.* 2015; **31**(15): 2586–2588.
PubMed Abstract | Publisher Full Text

Loayza-Puch F, Rooijers K, Buil LC, *et al.*: **Tumour-specific proline vulnerability uncovered by differential ribosome codon reading.** *Nature.* 2016; **530**(7591): 490–494.
PubMed Abstract | Publisher Full Text

Michel AM, Mullan JP, Velayudhan V, *et al.*: **RiboGalaxy: A browser based platform for the alignment, analysis and visualization of ribosome profiling data.** *RNA Biol.* 2016; **13**(3): 316–319.
PubMed Abstract | Publisher Full Text | Free Full Text

O'Connor P, Andreev D, Baranov P: **Surveying the relative impact of mRNA features on local ribosome profiling read density in 28 datasets.** *bioRxiv.* 2015; 018762.
Publisher Full Text

Popa A: **scriptWoolstenhulme_Defp2.R.** *Zenodo.* 2016.
Data Source

Popa A, Lebrigand K, Barbry P, *et al.*: **Pateamine A-sensitive ribosome profiling reveals the scope of translation in mouse embryonic stem cells.** *BMC Genomics.* 2016; **17**: 52.
PubMed Abstract | Publisher Full Text | Free Full Text

Schafer S, Adami E, Heinig M, *et al.*: **Translational regulation shapes the molecular landscape of complex disease phenotypes.** *Nat Commun.* 2015; **6**: 7200.
PubMed Abstract | Publisher Full Text | Free Full Text

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

Open Peer Review

Current Referee Status: ✓ ✓ ✓

Version 1

Referee Report 05 July 2016

doi:10.5256/f1000research.9644.r14283

Ghislain Bidaut
Integrative Bioinformatics, Inserm, Aix-Marseille Université, CNRS and Institut Paoli-Calmettes, Centre de Recherche en Cancérologie de Marseille, Marseille, France

I started my review by trying to reproduce the vignette example. I also made sure that all documentation was present in the reference package PDF, which was the case.

As stated in the paper, the Riboprofiling package has to be installed under R version 3.3 minimum. I installed the latest version at the date of review, which was 3.3.1+ BioC 3.3.

I started by the installation of the Riboprofiling package, which posed no problem.

I am not proficient with existing pipelines, and did not verify if Riboprofiling is more complete than them.

However, I found that an R implementation is nice since most users will be familiar with its installation, organisation and mode of use and won't have to find other tools elsewhere.

I have a few remarks on the installation and on the vignette example:

I had to install the Bioconductor Rsamtools which was not installed alongside Riboprofiling, while other dependencies were installed.

After running the covData <- riboSeqFromBAM(listInputBam, genomeName="hg19 »), I obtained the following warnings. Not sure if they impact the final result, but I nevertheless report them.

Warning messages:
1: In riboSeqFromBAM(listInputBam, genomeName = "hg19") :
paramScanBAM parameter is not a ScanBamParam object. Set to default NULL value!

2: In doTryCatch(return(expr), name, parentenv, handler) :
[knet_seek] SEEK_END is not supported for HTTP. Offset is unchanged.
3: In doTryCatch(return(expr), name, parentenv, handler) :
[bam_header_read] EOF marker is absent. The input is probably truncated.
4: In countShiftReads(exonGRanges[names(cdsPosTransc)], cdsPosTransc, :
Param motifSize should be an integer! Accepted values 3, 6 or 9. Default value is 3.
5: In doTryCatch(return(expr), name, parentenv, handler) :
[knet_seek] SEEK_END is not supported for HTTP. Offset is unchanged.

6: In doTryCatch(return(expr), name, parentenv, handler) :
[bam_header_read] EOF marker is absent. The input is probably truncated.
7: In countShiftReads(exonGRanges[names(cdsPosTransc)], cdsPosTransc, :
Param motifSize should be an integer! Accepted values 3, 6 or 9. Default value is 3.

In the vignette page 5, I would add the statement

library(GenomicAlignments)
before the readGAlignments call in order to make the code easier to work..

Vignete page 6. The instruction
txdb <- TxDb.Hsapiens.UCSC.hg19.knownGene
does not work. I do not have a TxDb.Hsapiens.UCSC.hg19.knownGene object in my session. I couldn't go further in running the provided tutorial. Could you clarify on how to obtain that object ?

Apart from the vignette, the paper itself is well written and can be followed simply through. An analysis case of ribosome stalling study is presented through a complete example and R code (analysis of Woolstenhulme et al dataset). The code works without problem until the invocation of library(BSgenome.Ecoli.NCBI.K12.MG1655). I couldn't get this package on the Bioconductor Web Site, and some clarification might be helpful for this step.

Minor remark:
The authors missed a couple of references in the last paragraph: (...) either using more specialisez riboseq tools such as (XXX, YYY) or directly within R.

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

Competing Interests: No competing interests were disclosed.

Referee Report 21 June 2016

doi:10.5256/f1000research.9644.r14282

Audrey M Michel
School of Biochemistry and Cell Biology, University College Cork, Cork, Ireland

RiboProfiling allows a number of ribo-seq specific analyses to be carried out using a single Bioconductor package in R and is supported by instructions in the corresponding reference and 'package' manuals. The automatic assignment of the P-site from either the 5' or 3' end of the ribosome footprint read is a useful feature. In addition the tool allows the quantification of the distribution of footprints across coding regions (CDS), 5' leader and 3' trailer regions at the metagene and individual gene transcript level. The authors use the reported stalling events in the Woolstenhulme *et al.,* 2015 [1] ribo-seq data to illustrate how RiboProfiling can be used to determine the ribosome occupancy on multiple codon motifs.

Overall, the RiboProfiling package allows some ribo-seq specific analyses to be automatically carried out using the RiboProfiling 'quick command' or by using the step-by-step commands. However I would argue that RiboProfiling is not more complete than existing tools. There is no assessment of triplet periodicity for example. The quality assessment carried out by RiboProfilig is the distribution of footprint read lengths

(Figure 1). While determining the read lengths in a ribo-seq dataset is a useful assessment, it is only one determinant of the quality of ribo-seq data. The RUST tool (O'Connor *et al.* 2015 [2]) as mentioned by the authors carries out normalisation of the data, but also allows for the quality of the ribo-seq and mRNA-seq data to be determined in terms of sequencing biases.

Nevertheless, RiboProfiling adds to the repertoire of publicly available ribo-seq tools which is useful to the community.

Questions:

1. In the manuscript the abbreviation TSS is used for Translation Start Site. However, on page 6 of the RiboProfiling manual “aroundPromoter: returns the genomic positions flanking the **transcript start site** (TSS) for the x% (3% default value) best expressed CDSs”. Should this be Translation start site?

2. It is not clear what the authors mean by tools such as (XXX, YYY) ' in the following sentence in the Summary section 'Following the overview of Ribo-seq experiments with ‘RiboProfiling, the output tables can then be easily integrated into more specialized dowstream analyses, either using more specialisez riboseq tools such as (XXX, YYY) or directly within R.' (There are also some typo errors in this sentence (dowstream; specialisez).

3. While following the instructions in the RiboProfiling manual, there were several operations for which I was missing the required library (e.g. library(Rsamtools), library(GenomicAlignments), library(TxDb.Hsapiens.UCSC.hg19.knownGene), etc). While it might be obvious which additional libraries are required after running the command, it may help users if they were specified in the manual along with library(RiboProfiling).

4. I ran the 'quick start' riboSeqFromBAM function on one of our own datasets and four plots were generated fine. However, the 'list of per ORF per codon coverage' was not output. Should this have been generated as stated in the manual?

5. What read lengths are considered in the RiboProfiling 'quick command' approach for the offset determination? On our own dataset the sumUp offset plot was generated showing an offset of -13 (which corresponds to our own estimation). However, this plot did not show the individual read lengths. Are read lengths sharing the same offset value used?

6. All of the step-by-step commands in the user manual ran fine on our own dataset and the corresponding plots were generated. However, there was an error when I tried to run the accompanying scriptWoolstenhulme_Defp2.R to generate the barplot in Figure 4B of the average ribosome occupancy on PP(X) motifs in the E.coli dataset (see further). So unfortunately I was unable to verify this part of the manuscript.

7. According to the manual, “consecutive motifs of 9 nucleotides (3 consecutive codons) overlap on 6 nucleotides. The Ribo-seq coverage is reported as the coverage on the 2nd codon in the motif considered as being in the P-site.” Is the average pause score in Figure 4B for ribosome occupancy of the second codon (P) or the last codon (X) or across the three codons?

8. The scriptWoolstenhulme_Defp2.R is hard-coded for the PPX motif and E.coli. I did not get to check if it can be easily adapted for other codon motifs in other datasets.

Below I provide a sub-set output of the scriptWoolstenhulme_Defp2.R with the error. If the sessionInfo() details can help troubleshoot the error, I can send it to the authors separately so as not clog up this report.
….
Import genomic features from the file as a GRanges object ... trying URL
'http://genomique.info/data/public/RiboProfiling/Escherichia_coli_str_k_12_substr_mg1655_no_rRNA_no

Content type 'unknown' length 2766800 bytes (2.6 MB)
==========
downloaded 2.6 MB

OK
Prepare the 'metadata' data frame ... OK
Make the TxDb object ... OK
[knet_seek] SEEK_END is not supported for HTTP. Offset is unchanged.
[bam_header_read] EOF marker is absent. The input is probably truncated.
[knet_seek] SEEK_END is not supported for HTTP. Offset is unchanged.
[bam_header_read] EOF marker is absent. The input is probably truncated.

Error in if (length(ixLengthPb)/length(testLength) >= 0.7) { :
  missing value where TRUE/FALSE needed
Calls: readStartCov -> readStartCov1Aln -> normRange
Execution halted

References

1. Woolstenhulme CJ, Guydosh NR, Green R, Buskirk AR: 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
2. O'Connor P: Surveying the relative impact of mRNA features on local ribosome profiling read density in 28 datasets. *bioRxiv*. 2015. Publisher Full Text

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

Competing Interests: No competing interests were disclosed.

Referee Report 20 June 2016

doi:10.5256/f1000research.9644.r14285

Olivier Namy[1], **Pierre Bertin**[2]
[1] Institute for Integrative Biology of the Cell (I2BC), University of Paris-Sud, Orsay, France
[2] Université Paris-Sud, Orsay, France

This article represents a useful package to analyze ribosome profiling data. We found that several important features of the Ribosome Profiling data processing are well implemented in this Bioconductor Package. Actually, the Kmer sizes distribution, the P-site location, the features (CDS, 5UTR, 3UTR) coverage distribution and the codon coverage are important and significant to explain Ribosome Profiling data and this package generates a great summary of these. However this package is not as complete as

suggested by the authors. Indeed an important point is the periodicity of footprints that is not addressed here. Moreover it could be interesting to add the ability to detect which reading phase is indeed read. This should be feasible since the P-site is well defined.

I include below a few comments for the authors.

- The algorithm of the readStartCov function is not sufficiently described. The only one sentence found, does not give enough information to understand how it works. For instance, "the read frequency distribution centered on the translation start site": "Centered" needs more explanation. This is important, because using our own datasets we were not able to obtain any convincing results about the position of the P-site. The lack of information prevented us from understanding why.

- In the Package Manual, when the aroundPromoter function is defined, the TSS is described as the "Transcript Start Site", which does not match with the "Translational Start Site" found in the article.

- The pipeline requires an annotation file with a specific format (TxDb object). Users have to be aware of that if they want to use their own genome.

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

Competing Interests: No competing interests were disclosed.