Published in final edited form as:

Curr Protoc Mol Biol. 2018 October ; 124(1): e66. doi:10.1002/cpmb.66.

Rfoot: Transcriptome-scale identification of RNA-protein complexes from ribosome profiling data

Zhe Ji*

Department of Pharmacology, Feinberg School of Medicine, Northwestern University, Chicago, IL 60611, USA. Department of Biomedical Engineering, McCormick School of Engineering, Northwestern University, Evanston, IL 60208, USA

Abstract

Ribosome profiling was developed to identify genome-wide RNA fragments associated with translating ribosomes. However, no experimental procedure was used to purify ribosome-RNA complexes. Actually, ribosome profiling is a transcriptomic RNase footprinting assay, which can identify both ribosomal and non-ribosomal protein-RNA complexes. Many sequencing reads represent functional RNA footprints associated with non-ribosomal proteins. Here I present a computational pipeline named Rfoot to systematically identify genome-wide non-ribosomal RNA footprints, based on highly localized read distribution feature. The analyses revealed native functional protein-RNA complexes in lncRNAs, 3'UTRs of mRNAs and all types of small noncoding RNAs. The computational tool is useful to discover novel noncoding functions of RNAs.

Keywords

RNase footprinting; Ribosome profiling; Non-ribosomal protein-RNA complex; lncRNA; small noncoding RNA

INTRODUCTION

Ribosome profiling was developed to map genome-wide ribosome-RNA complexes (Ingolia et al., 2013; Ingolia et al., 2009). However, during the experimental procedure, no step was used to specifically select ribosome-RNA complexes. Actually, ribosome profiling is a transcriptomic footprinting assay and can detect both ribosomal-RNA and non-ribosomal protein-RNA complexes (Ji et al., 2016). The reads representing non-ribosomal protein-RNA complexes tend to show highly localized distribution, while the reads representing ribosome complexes show 3-nt periodicity across actively translated open reading frames (ORFs). We used the Percentage of Maximum Entropy (PME) values to measure the uniformness of read distribution across a scanning window (See (Ji et al., 2016) for detailed explanation). Low PME values (close to 0) indicate that reads tend to show highly localized distribution, while

Contact information: Phone: (312) 503-5985, Fax: zhe.ji@northwestern.edu.

INTERNET RESOURCES

Rfoot software is available from <https://github.com/zhejilab/Rfoot>

high PME values (close to 1) indicate that the reads are uniformly distributed (Figure 1) (Ji et al., 2016). Based on these read distribution features, we developed two computational algorithms, including Rfoot to identify transcriptomic non-ribosomal protein-RNA complexes (Ji et al., 2016), and RibORF to characterize genome-wide translated ORFs (Ji et al., 2015).

In this protocol, I present the computational pipeline Rfoot to identify transcriptomic native non-ribosomal protein-RNA complexes using ribosome profiling datasets (Ji et al., 2015). The analyses revealed pervasive RNA footprints in 3'UTRs of mRNAs, lncRNAs and all types of small noncoding RNAs, such as snoRNAs, tRNAs, microRNAs and spliceosomal RNAs. Further characterization of the RNA footprints and associated protein complexes can lead to the discovery of novel noncoding functions of the RNAs.

BASIC PROTOCOL 1. Rfoot identifies non-ribosomal protein-RNA footprints

Following I will introduce steps of the Rfoot software. We published Rfoot in (Ji et al., 2016), and here I present an improved software version. I used a published ribosome profiling dataset in a human cell as the example. Similar analysis procedures can be used for other ribosome profiling datasets in different cells and model species.

Materials

Ribosome profiling datasets in Fastq format;
Genome assembly file in Fasta format;
Ribosomal RNA (rRNA) sequence file in Fasta format;
Transcript annotation file in genePred format;
Linux high performance computing cluster;
Perl program installation;
Read mapping software (such as Bowtie (Langmead and Salzberg, 2012) and Tophat (Kim et al., 2013))

Download Rfoot software

- 1 Download Rfoot package from <https://github.com/zhejilab/Rfoot/>.
Obtain the script "Rfoot.pl".

Map ribosome profiling reads to reference transcriptome and genome

- 2 Obtain ribosome profiling dataset with cycloheximide treatment or without drug treatment, trim 3' adapters of sequencing reads, map trimmed reads to rRNAs, and then align non-rRNA reads to the reference transcriptome and genome. Example commands are given below in steps 2a – 2f.
Example commands:

- 2a** Download an example ribosome profiling dataset from GEO database using the fastq-dump command line as a part of the NIH software sratoolkit, (<https://www.ncbi.nlm.nih.gov/sra/docs/toolkitsoft/>).

```
fastq-dump -Z SRR1802146 > SRR1802146.fastq
```

- 2b** Remove 3' adapters of sequencing reads, using "removeAdapter.pl" from <https://github.com/zhejilab/RibORF>.

```
perl removeAdapter.pl -f SRR1802146.fastq -a CTGTAGGCAC -o
adapter.SRR1802146.fastq
```

- 2c** Get human rRNA sequences from NCBI database, including 5S rRNA (NR_023363), 5.8S rRNA (NR_003285), 18S rRNA (NR_003286) and 28S rRNA (NR_003287). Put the rRNA sequences in the file "human.ribosomal.rna.fa" with fastq format.

- 2d** Use Bowtie to index rRNA sequences.

```
bowtie2-build human.ribosomal.rna.fa hg.ribosome
```

- 2e** Align trimmed ribosome profiling reads to rRNAs, and obtain non-rRNA reads.

```
bowtie2 -x hg.ribosome -U adapter.SRR1802146.fastq --un
norrna.adapter.SRR1802146.fastq -S
ribosome.adapter.SRR1802146.fastq
```

- 2f** Align non-rRNA reads to the reference transcriptome and genome, and obtain the alignment file in SAM format. The human reference transcriptome and genome can be obtained from the GENCODE database.

```
tophat --GTF gencode.v28.annotation.gtf --no-convert-bam -o
outputDir GRCh38genome.index norrna.adapter. SRR1802146.fastq
```

Obtain read alignment file in SAM format.

Run Rfoot program to identify transcriptomic RNase footprints

- 3** Run "Rfoot.pl" to identify transcriptomic non-ribosomal protein-RNA footprints.

Usage: perl Rfoot.pl -i readFile -t transcriptFile -o outputFile [-w windowSize] [-r readnumCutoff] [-f translatedORFFile]

- i readFile: input read mapping file in SAM format;

- t transcriptFile: transcripts annotation file in genePred format;
- o outputFile: output file reporting candidate non-ribosomal protein-RNA footprints;
- w windowSize [optional]: scanning window size, default: 60 nt;
- r readnumCutoff [optional]: cutoff of supported read number, default: 10;
- f translatedORFFile [optional]: file with translated ORF annotation in genePred format.

These regions will be excluded for the analyses of non-ribosomal protein-RNA complexes.

Example commands:

3a Obtain human reference transcriptome annotation:

```
wget
ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_28/
gencode.v28.annotation.gtf.gz
```

Get the gtfToGenePred command line from UCSC Genome Browser, and use the tool to convert the GTF file to genePred format:

```
wget
http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/gtfToGenePred
gtfToGenePred gencode.v28.annotation.gtf
gencode.v28.annotation.genePred.txt
```

3b Run “Rfoot.pl” to get the prediction:

```
perl Rfoot.pl -i SRR1802146.mapping.sam -t
gencode.v28.annotation.genePred.txt -o
candidate.nonribosome.sites.txt -f translatedORF.genePred.txt
```

Get the output file reporting candidate non-ribosomal protein-RNA footprints with the following columns: “transcript ID”, “chromosome”, “strand”, “footprint start site”, “footprint end site”, “footprint length”, “supported read number”, “position with maximum number of reads”, “maximum number of reads in a location”, “positions with supported reads (location:read number)”. Examples of non-ribosomal protein-RNA footprints are shown in Figure 2.

COMMENTARY

Background Information

Ribosome profiling was originally developed to identify ribosome protected RNA fragments (Ingolia et al., 2009; McGlincy and Ingolia, 2017). Cells were treated with translation inhibitors such as cycloheximide to pause ribosome elongation. After digesting unprotected RNA regions using RNase I, the cell lysate was subject to extensive centrifugation using a sucrose cushion. Protein-RNA complexes were enriched in cell pellets. RNA fragments were then isolated for next-generation sequencing. However, during ribosome profiling procedure, no experimental step was used to specifically select ribosome-RNA complexes, such as using ribosome antibodies. Based on the parameters of ultracentrifugation and sucrose cushion, we estimated that 7–10 S protein-RNA complexes will be suspended (Ji et al., 2016).

Indeed, we and others found that many ribosome profiling reads do not represent ribosome-RNA complexes (Ingolia et al., 2014; Ji et al., 2016). For example, many reads are from small noncoding RNAs, which are not translated (Ji et al., 2016). Careful examination of these reads showed that they represent functional RNA domains of small noncoding RNAs. Based on the observation, we repurposed ribosome profiling as a transcriptomic RNase footprinting assay, which can detect both ribosomal and non-ribosomal protein-RNA complexes (Ji et al., 2016).

There are 4 types of transcripts expressed in a mammalian cell, including mRNA, lncRNAs, pseudogenes and small noncoding RNAs. Untranslated regions of RNAs can have important biological functions, through functional RNA domains bound by protein complexes. For example, some lncRNAs play regulatory roles in various biological processes via functional RNA domains that regulate chromatin modifications, DNA transcription and mRNA stability (Rinn and Chang, 2012; Ulitsky and Bartel, 2013). Characterizing these functional RNA domains is critical for revealing novel biological roles of the RNAs.

Some protein complexes such as the 80S ribosome, can have the strong protection of RNA footprints from RNase I digestion. For other protein complexes transiently interacting with RNA domains such as the 40S scanning ribosomes, UV or formaldehyde is used to crosslink and stabilize protein-RNA complexes before RNase I digestion (Archer et al., 2016). The complexes after RNase I digestion were purified and associated RNA fragments were sequenced. In a CLIP-seq experiment, an antibody is used to isolate specific protein-RNA complexes of interest (Darnell, 2010). The RNase footprinting technologies can reveal novel functions of RNA molecules and associated proteins.

The non-ribosomal protein-RNA complexes in ribosome profiling datasets represent very stable complexes, which can protect RNA footprints from RNase I digestion like 80S ribosomes. These RNA footprints probably have important biological functions. Future research efforts characterizing these footprints and associated proteins can lead to discovering novel functions of the RNAs.

Critical Parameters

Distinguishing non-ribosomal protein-RNA footprints and short translated

ORFs—Users are commended to load SAM alignment file to a genome browser to visualize read distribution, and manually examine Rfoot predicted non-ribosomal protein-RNA footprints of interest. A source of false positives of non-ribosomal protein-RNA complexes is due to the reads from short translated ORFs (which can be 6 nt). Users can use the RibORF software (published in the same issue) to identify genome-wide translated ORFs, and exclude the translated regions using Rfoot optional parameter “-f translatedORFFile”. Many 5′UTRs of protein-coding genes contain multiple overlapping reading frames with unconventional start codons, which are difficult to be characterized. Users need to cautiously interpret localized reads in 5′UTRs. These are possibly from RNA translation.

Anticipated Results

Users will obtain a set of candidate non-ribosomal protein-RNA footprints across transcripts. These footprints can be helpful to examine novel functions of noncoding regions in RNAs.

Time Considerations

Depending on the sequencing depth and the number of annotated transcripts, Rfoot may take several hours to run.

Acknowledgments

I thank scientists who tried Rfoot and provided feedback to improve the software. The work was supported by grants to Z.J. from the National Institutes of Health (CA 207865) and from Northwestern University (the Searle Leadership Fund in the Life Sciences).

LITERATURE CITED

- Archer SK, Shirokikh NE, Beilharz TH, Preiss T. Dynamics of ribosome scanning and recycling revealed by translation complex profiling. *Nature*. 2016; 535:570–574. [PubMed: 27437580]
- Darnell RB. HITS-CLIP: panoramic views of protein-RNA regulation in living cells. *Wiley Interdiscip Rev RNA*. 2010; 1:266–286. [PubMed: 21935890]
- Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS. Genome-wide annotation and quantitation of translation by ribosome profiling. *Curr Protoc Mol Biol*. 2013; Chapter 4(Unit 4):18.
- Ingolia NT, Brar GA, Stern-Ginossar N, Harris MS, Talhouarne GJ, Jackson SE, Wills MR, Weissman JS. Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell reports*. 2014; 8:1365–1379. [PubMed: 25159147]
- Ingolia NT, Ghaemmighami S, Newman JR, Weissman JS. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*. 2009; 324:218–223. [PubMed: 19213877]
- Ji Z, Song R, Huang H, Regev A, Struhl K. Transcriptome-scale RNase-footprinting of RNA-protein complexes. *Nat Biotechnol*. 2016; 34:410–413. [PubMed: 26900662]
- Ji Z, Song R, Regev A, Struhl K. Many lncRNAs, 5′UTRs, and pseudogenes are translated and some are likely to express functional proteins. *Elife*. 2015; 4:e08890. [PubMed: 26687005]
- Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*. 2013; 14:R36. [PubMed: 23618408]
- Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nature methods*. 2012; 9:357–359. [PubMed: 22388286]

- McGlinchy NJ, Ingolia NT. Transcriptome-wide measurement of translation by ribosome profiling. *Methods*. 2017; 126:112–129. [PubMed: 28579404]
- Rinn JL, Chang HY. Genome regulation by long noncoding RNAs. *Annu Rev Biochem*. 2012; 81:145–166. [PubMed: 22663078]
- Ulitsky I, Bartel DP. lincRNAs: genomics, evolution, and mechanisms. *Cell*. 2013; 154:26–46. [PubMed: 23827673]

KEY REFERENCE

- The Rfoot algorithm was originally published in: Ji Z, Song R, Huang H, Regev A, Struhl K. Transcriptome-scale RNase-footprinting of RNA-protein complexes. *Nat Biotechnol*. 2016; 34:410–413. [PubMed: 26900662]
- The RibORF program presented in the following paper is complementary and is to identify translated ORFs: Ji Z, Song R, Regev A, Struhl K. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *Elife*. 2015; 4:e08890. [PubMed: 26687005]

Figure 1. Distinguishing reads representing non-ribosomal protein-RNA complexes and ribosome complexes. Read distribution patterns representing non-ribosomal protein binding (with low PME values) and translated ORFs (high PME values).

Figure 2.
Examples of non-ribosomal protein-RNA footprints in a small noncoding RNA RNU11 and a lncRNA NEAT1.