Application Note

uORF-Tools – Workflow for the determination of translation-regulatory upstream open reading frames

Anica Scholz[1,†], Florian Eggenhofer[2,†,*], Rick Gelhausen[2], Björn Grüning[2], Kathi Zarnack[3], Bernhard Brüne[1], Rolf Backofen[2,4] and Tobias Schmid[1,*]

[1] Institute of Biochemistry I, Faculty of Medicine, Goethe-University Frankfurt, Theodor-Stern-Kai 7, 60590 Frankfurt am Main, Germany,
[2] Bioinformatics Group, Department of Computer Science, University of Freiburg, Georges-Köhler-Allee, 79110 Freiburg, Germany,
[3] Buchmann Institute for Molecular Life Sciences (BMLS), Goethe-University Frankfurt, Max-von-Laue-Straße 15, 60438 Frankfurt am Main, Germany,
[4] Centre for Biological Signalling Studies (BIOSS), University of Freiburg, 79110 Freiburg, Germany.

[†] Authors contributed equally.
* Corresponding authors (egg@bioinformatik.uni-frankfurt.de, t.schmid@biochem.uni-frankfurt.de)

Keywords:
Translation, Ribosome profiling, Gene regulation, Next-generation sequencing, RNA, Snakemake

Abstract

Ribosome profiling (ribo-seq) provides a means to analyze active translation by determining ribosome occupancy in a transcriptome-wide manner. The vast majority of ribosome protected fragments resides within the protein-coding sequence of mRNAs. However, commonly reads are also found within the transcript leader sequence (TLS) (aka 5' untranslated region) preceding the main open reading frame (ORF), which indicates the translation of regulatory upstream ORFs (uORFs). Here, we present a workflow for the identification of functional uORFs, which contribute to the translational regulation of their associated main ORFs. The workflow is available as free and open Snakemake workflow. Furthermore, we provide a comprehensive human uORF annotation file, which can be used within the pipeline, thus reducing the runtime. (Availability: https://github.com/anibunny12/uORF-Tools)

Introduction

The recently developed ribosome profiling (ribo-seq) technology provides an elegant means to determine actively translated regions across the entire transcriptome by selective sequencing of ribosome-protected fragments (RPF) (Ingolia *et al.*, 2009). In line, reads in ribo-seq analyses are predominantly annotated to the protein-coding regions. Interestingly though, while the 3' untranslated regions (UTRs) of transcripts usually lack RPFs, the latter are commonly observed in the 5'UTRs. Such actively translated regions are indicators for the presence of upstream open reading frames (uORFs), which represent short, peptide-coding sequences containing both start and stop codons in frame. Consequently, 5' UTRs are also referred to as transcript leader sequences (TLS) (Wethmar, 2014). uORFs are considered to influence the translation of the associated main ORFs, i.e. efficient translation of the uORF commonly restricts translation initiation at the start site of the main ORF (Somers *et al.*, 2013). Various tools determine uORFs based on sequence features within the TLS (McGillivray *et al.*, 2018), and only recently approaches have been developed to use experimental ribo-seq data to identify actively translated uORFs (Calviello *et al.*, 2015; Zhang *et al.*, 2017).

With the present workflow, we aim to provide a pipeline that allows for the identification of those uORFs that are not only translated, but further regulate the translation of the associated main ORFs. Using ribo-some profiling data (ribo-seq and associated RNA-seq) in a fastq format, and appropriate genome and annotation files as input, *uORF-Tools* determines the experiment-specific translation-regulatory uORFs. While a uORF annotation file can be generated for individual experiments, a comprehensive human uORF annotation file is provided, which is based on data sets from nine human ribosome profiling data series (see Supplementary Table 1). uORF-Tools is provided as free and open workflow, available at: https://github.com/anibunny12/uORF-Tools.

Implementation and workflow

The workflow is based on *Snakemake* (Köster and Rahmann, 2012) and automatically installs all tool dependencies via *bioconda* (Grüning *et al.*, 2018). Ribosome profiling data sets containing ribo-seq and corresponding RNA-seq data, provided as raw, compressed fastq files, are required as input for uORF-Tools (Figure 1). Initial processing of the raw reads includes adapter and quality trimming using *Trim galore!* (Martin, 2011), and rRNA removal using *sortMeRNA* (Kopylova *et al.*, 2012). A quality assessment report is generated with *FastQC* (Andrews, 2014) for each of these steps. To allow for mapping using *STAR* (Dobin *et al.*, 2013), a genome fasta and an annotation gtf file (e.g. from Gencode) have to be supplied by the user. The mapped reads then enable the identification of translated uORFs using *Ribo-TISH* (Zhang *et al.*, 2017). The experiment-specific uORF annotation file generated within the workflow or the comprehensive human uORF annotation file, which is supplied with *uORF-Tools*, are then used for the subsequent steps. To determine translational regulation in general, it is important

to calculate differential translational efficiencies (TEs). The R package *Xtail* (Xiao *et al.*, 2016) is designed to determine TEs from ribosome profiling experiments. To this end, *Xtail* calculates size factors using *DESeq2* (Love *et al.*, 2014), based on which normalized expression values are calculated. The resulting changes in RPF-to-mRNA ratios are eventually evaluated using a tailored statistical model. However, the calculation of size factors for library subsets, such as uORFs, is not reliable, since the number of reads is too low to allow for the use of the median size factor approach of *DESeq2*. Thus, we applied *Xtail* for both uORFs and main ORFs, with size factors determined *a priori* with *DESeq2* for the entire ribosome profiling libraries, thereby also limiting the influence of stimulus-dependent changes in the sizes of selected library subsets (details see Supplementary Methods).

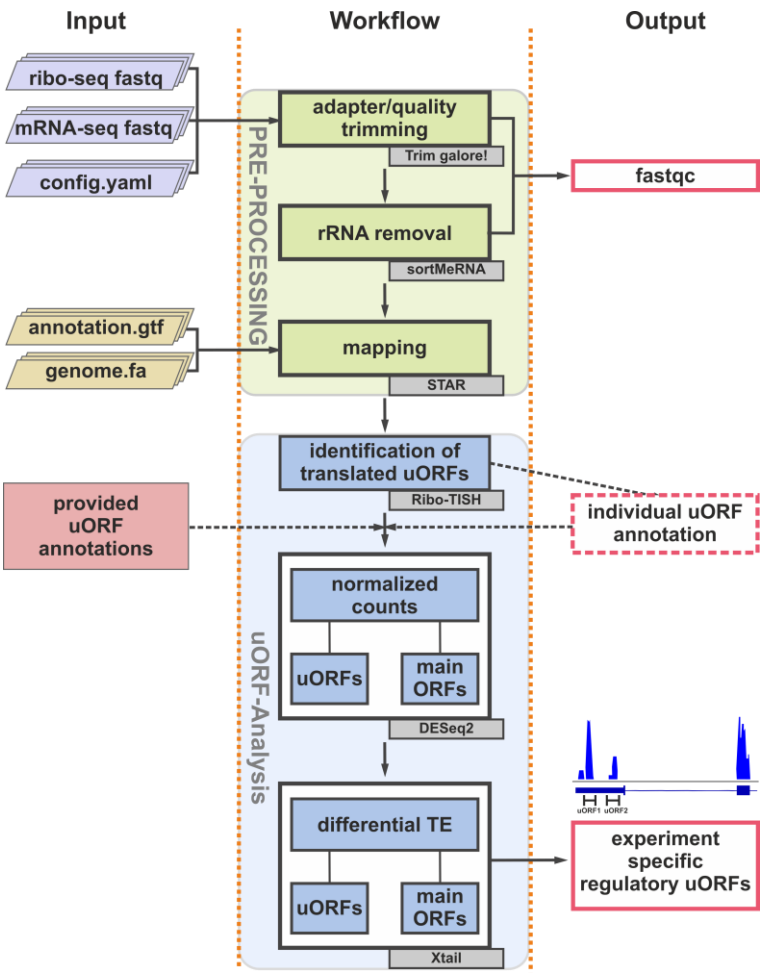

Fig. 1. *uORF-Tools* – Workflow for the determination of translation-regulatory uORFs. Required input is shown on the left, a simplified depiction of processing in the center, and results on the right.

The final output of *uORF-Tools* is an easy to interpret list containing the TEs of the uORFs and their associated main ORFs, highlighting if both are regulated in the same direction or reciprocally under the experiment-specific conditions. Moreover, the workflow generates an interactive summary report and tracks the fraction of initial reads used in each step.

Results

uORF-Tools is provided as a readily deployable and well documented *Snakemake* workflow. The input, processing steps, and used tools are clearly defined and enable reproducible analyses. The workflow provides a processing pipeline for ribosome profiling data as well as a *de novo* annotation of actively translated uORFs. Furthermore, it identifies changes in translation efficiency of uORFs and their associated main ORFs, thus allowing for the identification of translation-regulatory uORFs (Supplementary Fig. S1). In addition, we provide a comprehensive human uORF annotation file (bed file format) containing 3873 uORFs, identified in a total of 22 experimental data sets from nine data series (Supplementary Table 1). The compilation of the uORFs identified in different data sets suggests that addition of further data sets to the analysis will only marginally increase the number of uORFs for the comprehensive annotation file (Supplementary Fig. S2). *uORF-Tools* generates an intuitive, easy to interpret output file, allowing for the identification of translation-regulatory uORFs that are active under specific experimental conditions. As a side note, since *uORF-Tools* inherently provides changes in translational efficiencies for the main ORFs, it can also be used to assess translational changes in general.

Funding

This work was supported by the German Research Foundation (DFG) SCHM 2663/3 and by the High Performance and Cloud Computing Group, University Tübingen via bwHPC, DFG INST 37/935-1 FUGG. A.S. was supported by the MainCampus doctus program of the Stiftung Polytechnische Gesellschaft Frank-furt, R.G. by DFG grant BA 2168/21-1 SPP 2002 Small Proteins in Prokaryotes.

References

Andrews, S. (2010) FastQC a quality control tool for high throughput sequence data. www.bioinformatics.babraham.ac.uk/projects/fastqc.

Calviello, L. *et al.* (2016) Detecting actively translated open reading frames in ribosome profiling data. *Nat. Methods*, **13**, 165-170.

Dobin, A. *et al.* (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.

Grüning, B. *et al.* (2018) Bioconda: Sustainable and comprehensive software distribution for the life sciences. *Nat. Methods*, **15**, 475-476.

Ingolia, N.T. *et al.* (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, **324**, 218-223.

Kopylova, E. *et al.* (2012) SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics*, **28**, 3211–3217.

Köster, J. and Rahmann, S. (2012) Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics*, **28**, 2520-2522.

Love, M.I. *et al.* (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17, 10. (https://doi.org/10.14806/ej.17.1.200)

McGillivray, P. *et al.* (2018) A comprehensive catalog of predicted functional upstream open reading frames in humans. *Nucleic Acids Res.*, **46**, 3326-3338

Somers, J. *et al.* (2013) A perspective on mammalian upstream open reading frame function. *Int. J. Biochem. Cell Biol.*, **45**, 1690-700.

Wethmar, K. (2014) The regulatory potential of upstream open reading frames in eukaryotic gene expression. *Wiley Interdiscip. Rev. RNA*, **5**, 765-778.

Xiao, Z. *et al.* (2016). Genome-wide assessment of differential translations with ribosome profiling data. *Nat. Commun.*, **7**, 11194.

Zhang, P. *et al.* (2017) Genome-wide identification and differential analysis of translational initiation. *Nat. Commun.*, **8**, 1749.