

HRIBO - High-throughput analysis of bacterial ribosome profiling data

Rick Gelhausen^{1,*}, Florian Heyl¹, Sarah L. Svensson², Kathrin Froschauer², Lydia Hadjeras², Cynthia M. Sharma², Florian Eggenhofer^{†,1}, Rolf Backofen^{1,3,4,†,*}

1 Bioinformatics Group, Department of Computer Science, University of Freiburg, Georges-Koehler-Allee 106, 79110 Freiburg, Germany

2 Chair of Molecular Infection Biology II, Institute of Molecular Infection Biology (IMIB), Josef-Schneider-Str. 2 / Bau D15, University of Würzburg, Germany

3 ZBSA Centre for Biological Systems Analysis, University of Freiburg, Hauptstr. 1, 79104 Freiburg, Germany

4 Signalling Research Centres BIOS and CIBSS, University of Freiburg, Schanzlestr. 18, 79104 Freiburg, Germany

* Corresponding authors: gelhausr@informatik.uni-freiburg.de, rolf@informatik.uni-freiburg.de

† These authors contributed equally to this work

Abstract

Motivation: Ribosome profiling (Ribo-seq) is a powerful approach based on ribosome-protected RNA fragments to explore the transcriptome of a cell, and is especially useful for the detection of small proteins (≤ 70 amino acids) that are recalcitrant to biochemical and *in silico* approaches. While pipelines are available to analyze Ribo-seq data, none are designed explicitly for the analysis of Ribo-seq data from prokaryotes, nor are they focused on the discovery of unannotated open reading frames (ORFs) in bacteria.

Results: We present HRIBO (High-throughput annotation by Ribo-seq), a workflow to enable reproducible and high-throughput analysis of bacterial Ribo-seq data. The workflow performs all required pre-processing and quality control steps. Importantly, HRIBO outputs annotation-independent ORF predictions based on two complementary bacteria-focused tools, and integrates them with additional features. This facilitates the rapid discovery of novel ORFs and their prioritization for functional characterization.

Availability: HRIBO is a free and open source project available under the GPL-3 license at: <https://github.com/RickGelhausen/HRIBO>

Introduction

Ribosome profiling (Ribo-seq) is an RNA-seq-based approach that identifies the ribosome-bound fraction of the transcriptome as a proxy for protein expression (Fig 1). Because RNase digestion creates so-called ribosome footprints, it also allows identification of open-reading frame (ORF) boundaries. This makes it remarkably suited to detect small proteins, which are currently underrepresented in genome annotations due to their unique features that preclude conventional biochemical or *in silico* detection [14]. A parallel whole-transcriptome library allows calculation of translation efficiency (TE, the ratio of footprint coverage to transcriptome coverage) and identification of differentially-expressed ORFs. There are existing workflows for the analysis of eukaryotic Ribo-seq data [2], but a specialized solution for bacteria is still missing. Here, we present HRIBO

(High-throughput annotation by Ribo-seq), a computational pipeline that processes and analyses data from any bacterial Ribo-seq experiment, but also detects translated novel ORFs. This tool is compatible with bacterial annotation and circular genomes, uses an optimized mapping approach suitable for small bacterial genomes, integrates machine learning-based ORF prediction tools designed for/trained on bacterial ORF features and also include two differential expression tools designed for Ribo-seq data. We implemented HRIBO based on snakemake [7], which allows highly reproducible and fully automatic data analysis.

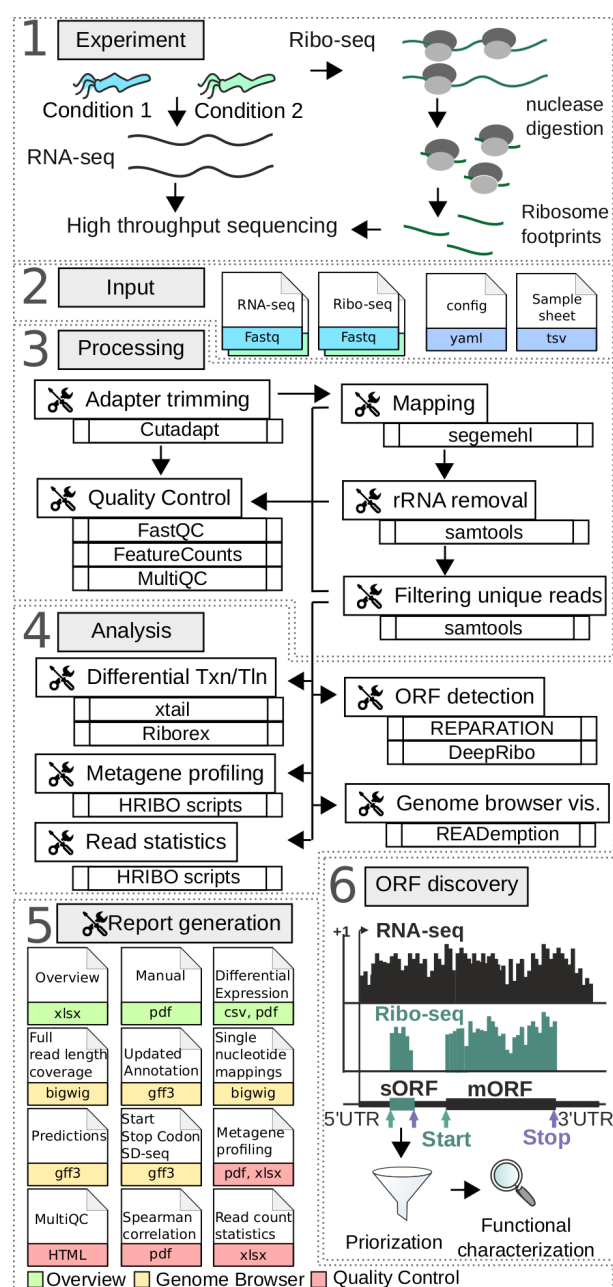


Figure 1. Outline of a bacterial Ribo-seq experiment and data analysis by the HRIBO pipeline, (input, processing analysis, output). For details see: <https://hribo.readthedocs.io/en/latest>.

Approach

HRIBO automatically retrieves tools from bioconda [6] and performs all necessary steps of processing, with pinned and tested tool versions. Detailed documentation, with examples, is available at:

<https://hribo.readthedocs.io/en/latest/>

HRIBO (Fig 1, Input) requires the sequencing data files (paired RNA-seq and Ribo-seq libraries for each sample), the genomic sequence/annotation of the organism, and a sample sheet that captures the experimental setup by associating samples/replicates with their experimental condition. As a first step (Fig 1, Processing), HRIBO performs adapter trimming with **Cutadapt** [10] and then maps the reads to the genome with **segemehl**, which has higher sensitivity than other mappers [12], but its high computational costs are still acceptable for small genomes.

Multimapping/rRNA reads are then removed before further processing. **FastQC** [1] and **featurecount** [9] reports are created after each processing step and aggregated with **MultiQC** [4], enabling the investigator to identify problems with either the experimental setup (e.g. insufficient rRNA depletion), or the pre-processing (e.g. untrimmed adapters). The resulting mapped reads are then used in subsequent steps (Fig 1, Analysis) to calculate read statistics, detect expressed ORFs (annotated or novel), compute differential expression and coverage tracks, and to perform metagene profiling of global ribosome density at start codons. We have included two complementary tools, both specifically developed for bacterial organisms, to detect ORFs. **REPARATION** [11] uses a random forest-based machine-learning approach. **DeepRibo** [3] is based on convolutional and recurrent neural network approaches. The results of the two prediction tools are aggregated by newly-developed scripts and compiled into a list of detected ORFs. Novel and annotated ORFs that are differentially expressed (transcription and translation) between samples/conditions are detected using the complementary Ribo-seq-specific tools **xtail** [16] and **Riborex** [8]. The ORF list is then enriched with additional computed context information for each ORF, such as length, sequence, normalized read counts, TE, available annotation/novelty, differential expression and which conditions/replicate the ORF was detected in (see Table 1).

Table 1. Result overview table, showing the first ten open reading frame entries for the example *Salmonella typhimurium* data set [15] analysed with HRIBO. The table contains both annotated and predicted ORFs, from **DeepRibo** and **Reparation** and with context information like length, nucleotide sequence, amino acid sequence, normalized read counts. Moreover the differential transcription and translation is analyzed, with **Riborex** and **xtail**, for all pairwise combinations of provided experimental conditions and added to the table. The evidence column highlights in which experimental condition the ORF was detected by **DeepRibo** and/or **Reparation**. The table, by integrating a plethora of result information enables a stringent screen for candidates for further experimental analysis. For the full output please refer to the complete analyzed dataset referenced in the text.

	Genome	Start	Stop	Strand
1	NC_016855.1	94	378	+
2	NC_016855.1	391	546	+
3	NC_016855.1	668	841	+
4	NC_016855.1	712	1154	+
5	NC_016855.1	1053	1265	-
6	NC_016855.1	1333	1599	+
7	NC_016855.1	1354	1599	+
8	NC_016855.1	1722	1844	+
9	NC_016855.1	1849	1926	+
10	NC_016855.1	1907	2788	+
	Locus_tag	Name	Gene_name	Length
1	STM14_RS00005	WP_000477658.1	STM14_RS00005	285
2	STM14_RS00010	WP_000983042.1	STM14_RS00010	156
3				174
4	STM14_RS00015		STM14_RS00015	443
5	STM14_RS25775		STM14_RS25775	213
6	STM14_RS00020	WP_010999936.1	STM14_RS00020	267
7				246

Continued on next page

Table 1 – continued from previous page

8	STM14_RS00025	WP_010999937.1	STM14_RS00025	123
9	STM14_RS00030	WP_001541552.1	STM14_RS00030	78
10	STM14_RS00035	WP_001738692.1	STM14_RS00035	882
	Codon_count	Start_codon	Stop_codon	Nucleotide_seq
1	95	GTG	TAG	GTGTTTAAAGT...
2	52	ATG	TGA	ATGATGCCCG...
3	58	ATG	TGA	ATGAGACTCA...
4	147.666666666667	GTG	TAG	GTGTTTATCC...
5	71	GTG	GCT	GTGTTGGCGC...
6	89	TTG	TGA	TTGATCCTTG...
7	82	ATG	TGA	ATGTCGCAGA...
8	41	GTG	TAG	GTGCTACCC...
9	26	ATG	TGA	ATGCTCAGAA...
10	294	GTG	TAA	GTGCAGGCAG...
	Aminoacid_seq	RIBO-WT-1_TE	RIBO-csrA-1_TE	RIBO-WT-1_rpk
1	VFKYWPTFVQ...	1.03	1.28	8333.98
2	MMPVKEATEK...	0.23	0.53	3349.62
3	MRLRKGHGLT...	3.53	4.16	27300.97
4		2	2.25	9114.7
5	VLAHGLRFRI...	0.08	0.05	446.04
6	LILEEAEMSQ...	1.54	2.57	145357.71
7	MSQIENAVTS...	1.63	2.78	156607.67
8	VSTRDQKSKN...	0.47	0.08	39779.51
9	MLRKFAQYLF...	1.05	1.27	96225.41
10	VQAEVTDTRT...	0.38	0.41	9963.92
	RIBO-csrA-1_rpk	RNA-WT-1_rpk	RNA-csrA-1_rpk	Reparation_probability
1	7790.51	8062.07	6106.8	0
2	5235	14728.79	9950.54	0
3	26987.32	7729.83	6488.12	0
4	9332.6	4554.14	4141.12	0
5	239.63	5262.09	5300.16	0
6	168225.81	94661.43	65537.35	0
7	181030.42	96135.8	65013.12	0.89403532155948
8	8091.9	84289.68	99814.25	0
9	96847.48	91246.64	75985.92	0
10	10300.84	26241.54	25066.08	0
	Deepribo_rank	Deepribo_score	riborex_pvalue	riborex_pvalue_adjusted
1	0	0	0.941477067882842	0.999989307718023
2	0	0	0.597391183938145	0.999989307718023
3	1892	1.6650712490081787	0.997171419377332	0.999989307718023
4	0	0	0.960967679360969	0.999989307718023
5	0	0	0.726070473492159	0.999989307718023
6	0	0	0.637030257270868	0.999989307718023
7	1329	2.27506685256958	0.634657235318253	0.999989307718023
8	0	0	0.217621786810577	0.999989307718023
9	0	0	0.958824566490529	0.999989307718023
10	0	0	0.942768501450736	0.999989307718023
	riborex_log2FC	xtail_pvalue	xtail_pvalue_adjusted	xtail_log2FC
1	-0.07341663376001	0.873062298756547	0.999974167017596	0.078511666975927
2	-0.979936528807948	0.270196556326851	0.999974167017596	1.01144416971949
3	-0.00593280091088	0.979199162413816	0.999974167017596	0.008286175420051
4	0.058788293373097	0.912209424782855	0.999974167017596	-0.049445595388108
5	1.13683834544813	0.8948362562224	0.999974167017596	-1.09960664072743
6	-0.511218542276689	0.005398874780438	0.255060844852764	0.512632799154282
7	-0.543386834447704	0.00349134620398	0.19966547819366	0.544596384566803
8	2.77139728797676	4.95877353119331E-09	1.36636750664063E-06	-2.77852218807137
9	-0.043308518396802	0.878349492318766	0.999974167017596	0.04491302875467
10	0.115951404390933	0.68173234285987	0.999974167017596	-0.114182260430041
	Evidence			
1				
2				
3	deepribo-WT-1 deepribo-csrA-1			
4				
5				
6				
7	deepribo-WT-1 deepribo-csrA-1 reparation-WT-1 reparation-csrA-1			
8				
9				
10				

Interesting candidates can then be inspected by either full-coverage [5] or single nucleotide mapping (5'/3' end or centre of read) genome browser tracks. Moreover, we developed metagene profiling scripts that analyze a range of read lengths for read density around the start codon (see Figure 2). Finally, the results are collected into a report (Fig 1, Report) that can be easily distributed and contains a detailed manual. An example report for a published dataset [13] can be found here:

ftp://biftp.informatik.uni-freiburg.de/pub/HRIBO/HRIBO1.4.0_17-04-20.zip

Conclusion

HRIBO is a reproducible and standardized pipeline to process Ribo-seq data-sets from bacterial organisms, from pre-processing and quality control, to ORF prediction and differential expression analyses. Read information and additional computed features for both annotated and predicted ORFs are summarized, allowing their prioritization for characterization and genome browser tracks for manual inspection. This streamlines the

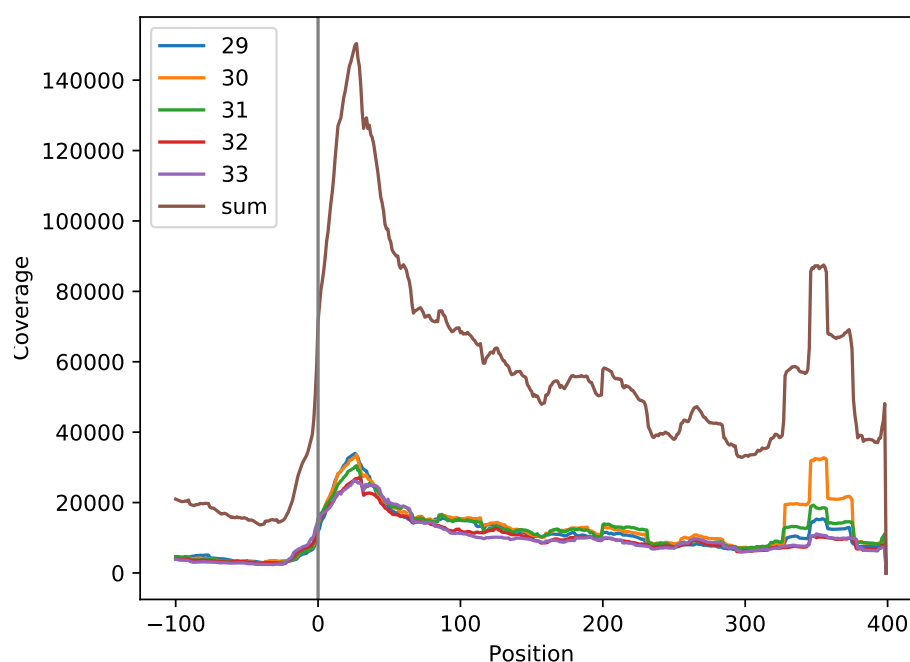


Figure 2. Fragment length specific meta gene profiling from analyzing the example *Salmonella typhimurium* data set with HRIBO version 1.4.0. Metagene profiling is a quality check for Ribo-seq experiments showing the average read distribution around the start codon.

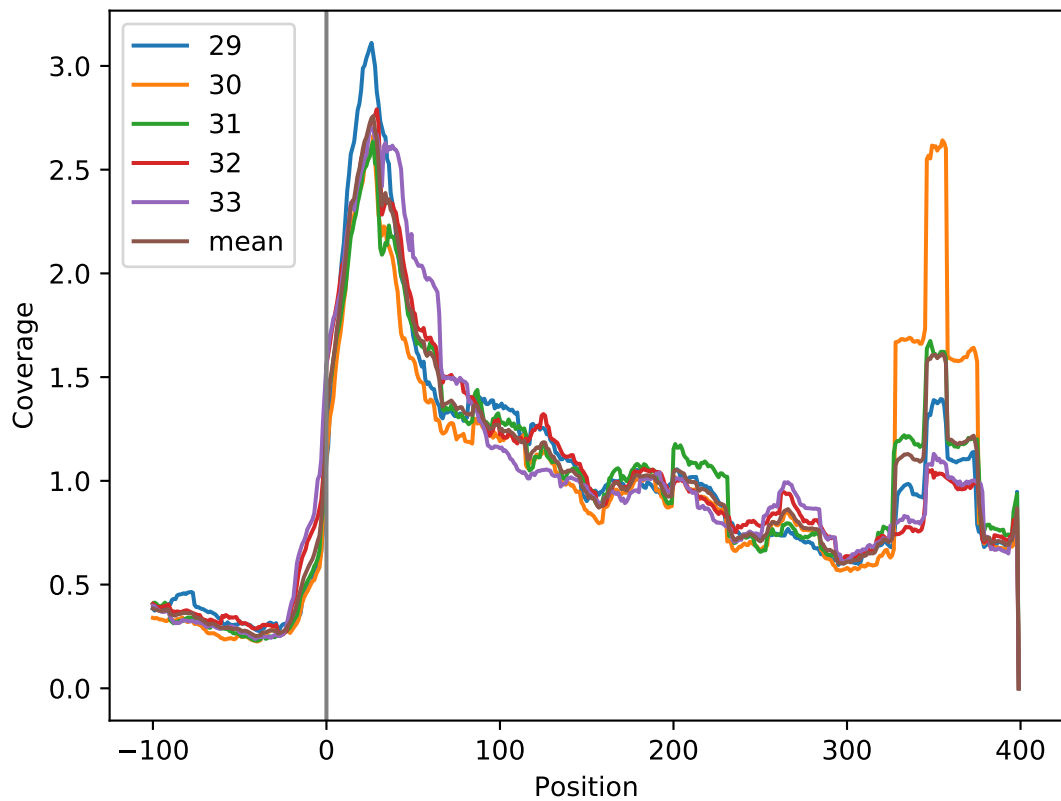
selection of high-confidence, functionally important, translated novel ORFs for further experimental investigation. We plan to add an analysis of translation initiation site data [15] in the future.

Funding: This work was supported by the German Research Foundation (DFG) SCHM 2663/3; the High Performance and Cloud Computing Group, University Tübingen via bwHPC; DFG INST 37/935-1 FUGG. R.G.; DFG grant BA 2168/21-1 SPP 2002; DFG grant 322977937/GRK2344 2017 MeInBio – BioInMe Research Training Group; BMBF ”Verbundprojekt RNAProNet, Integration des Netzwerks aus RNA- und Proteinbasierter Regulation - Teilprojekt B - 031L0164B”

References

1. Simon Andrews et al. Fastqc: a quality control tool for high throughput sequence data, 2010.
2. Betty Y Chung et al. The use of duplex-specific nuclease in ribosome profiling and a user-friendly software package for ribo-seq data analysis. *Rna*, 21(10):1731–1745, 2015.
3. Jim Clauwaert et al. Deepribo: a neural network for precise gene annotation of prokaryotes by combining ribosome profiling signal and binding site patterns. *Nucleic acids research*, 47(6):e36–e36, 2019.
4. Philip Ewels et al. Multiqc: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19):3047–3048, 2016.
5. Konrad U Förstner et al. Reademption—a tool for the computational analysis of deep-sequencing-based transcriptome data. *Bioinformatics*, 30(23):3421–3423, 2014.

6. Björn Grüning et al. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature methods*, 15(7):475–476, 2018.
7. Johannes Köster et al. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, 2012.
8. Wenzheng Li et al. Riborex: fast and flexible identification of differential translation from ribo-seq data. *Bioinformatics*, 33(11):1735–1737, 2017.
9. Yang Liao et al. featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, 2014.
10. Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17(1):10–12, 2011.
11. Elvis Ndah et al. Reparation: ribosome profiling assisted (re-) annotation of bacterial genomes. *Nucleic acids research*, 45(20):e168–e168, 2017.
12. Christian Otto et al. Lacking alignments? the next-generation sequencing mapper segemehl revisited. *Bioinformatics*, 30(13):1837–1843, 2014.
13. Anastasia H Potts et al. Role of csra in stress responses and metabolism important for salmonella virulence revealed by integrated transcriptomics. *PloS one*, 14(1), 2019.
14. Gisela Storz et al. Small proteins can no longer be ignored. *AnnuRevBiochem*, 83:753–77, 2014.
15. Jeremy Weaver et al. Identifying small proteins by ribosome profiling with stalled initiation complexes. *mBio*, 10(2):e02819–18, 2019.
16. Peng Zhang et al. Genome-wide identification and differential analysis of translational initiation. *Nature communications*, 8(1):1–14, 2017.



[illegible]