riboseq-flow: a streamlined, reliable pipeline for ribosome profiling data analysis and quality control

Ira A. Iosub[1,2,3], Oscar G. Wilkins[1,4], Jernej Ule[1,2,3]

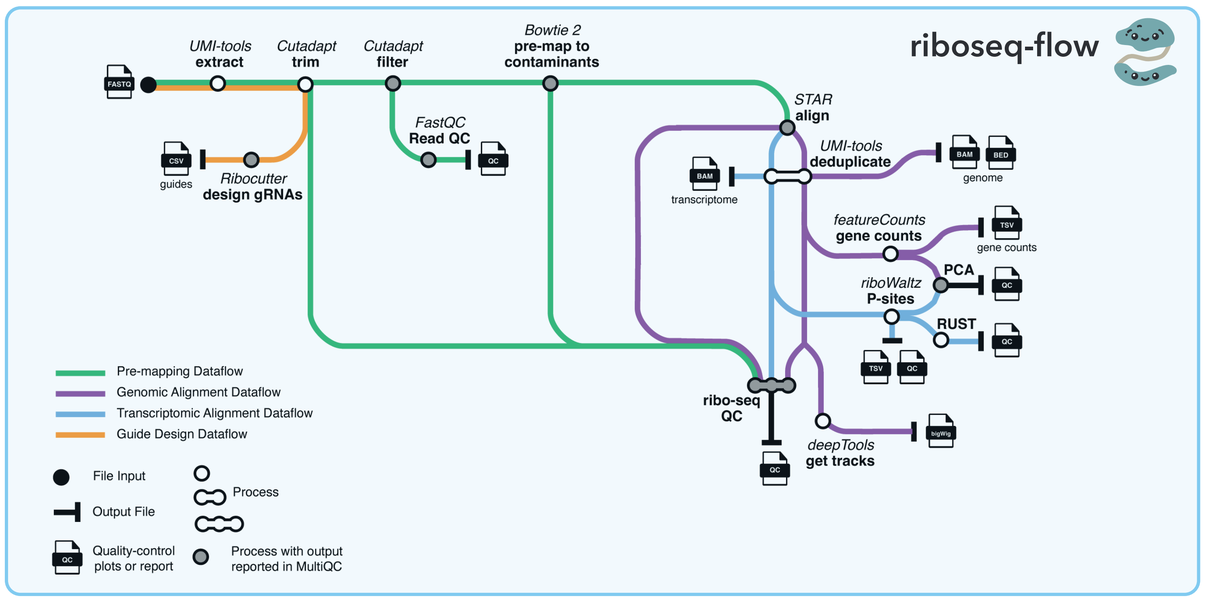
1 The Francis Crick Institute, London, UK
2 UK Dementia Research Institute at King's College London, London, UK
3 Department of Basic and Clinical Neuroscience, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK
4 Department of Neuromuscular Diseases, UCL Queen Square Institute of Neurology, UCL, London, UK

Summary

Ribosome profiling (Ribo-seq) is a powerful technique to study translation at a transcriptome-wide level. However, ensuring good data quality is paramount for accurate interpretation, as is ensuring that the analyses are reproducible. We introduce a new Nextflow[1,2] DSL2 pipeline, riboseq-flow[3], designed for processing and comprehensive quality control (QC) of ribosome profiling experiments. Riboseq-flow is user-friendly, versatile and upholds high standards in reproducibility, scalability, portability, and features version control and continuous integration. It enables users to efficiently analyse multiple samples in parallel and helps them evaluate the quality and utility of their data based on the detailed metrics and visualisations that are automatically generated. Here we present the primary riboseq-flow data processing steps and a selection of example outputs using public ribosome profiling data from human cells[4].

- ✓ Reproducible
- ✓ Portable
- ✓ Scalable

github.com/iraiosub/riboseq-flow



Outputs

Per-sample Ribo-seq QC with read-length stratified metrics

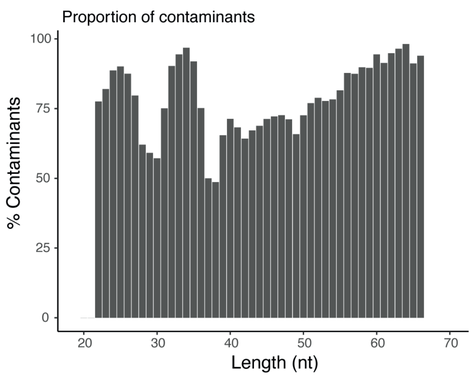


Fig. 1 Percentage of reads mapping to abundant non-coding RNA contaminants such as rRNA.

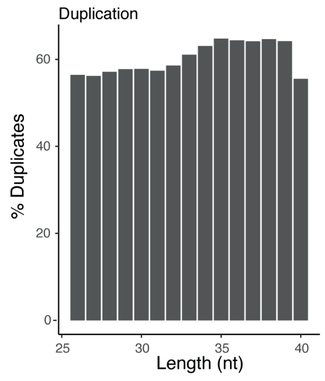


Fig. 2 Duplication percentage for reads of expected length.

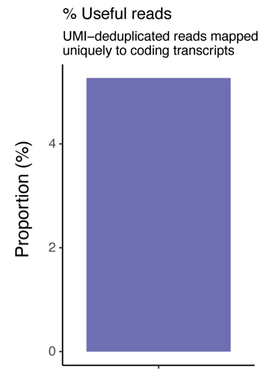


Fig. 3 Percentage of useful reads: ratio of reads mapped uniquely to protein-coding transcripts to the total number of adaptor- trimmed input reads.

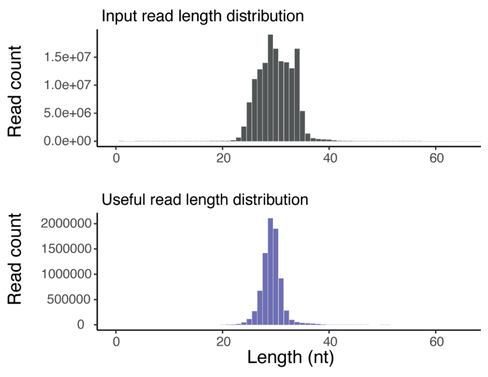


Fig. 4 Read length distribution of adaptor-trimmed input reads (top) and useful reads (mapped uniquely to protein-coding transcripts).

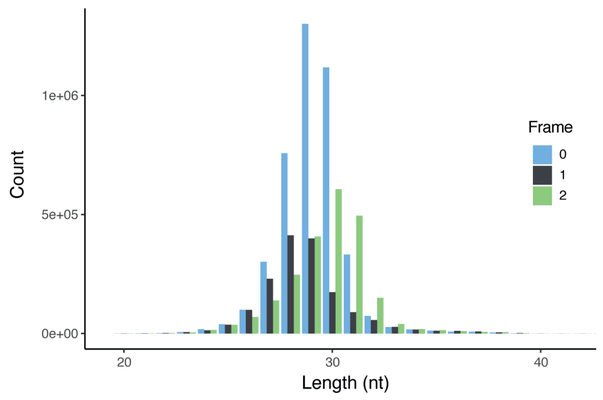


Fig. 5 Triplet periodicity plot showing the distribution of sub-codon positions of 5’ ends of RPFs to the protein coding regions.

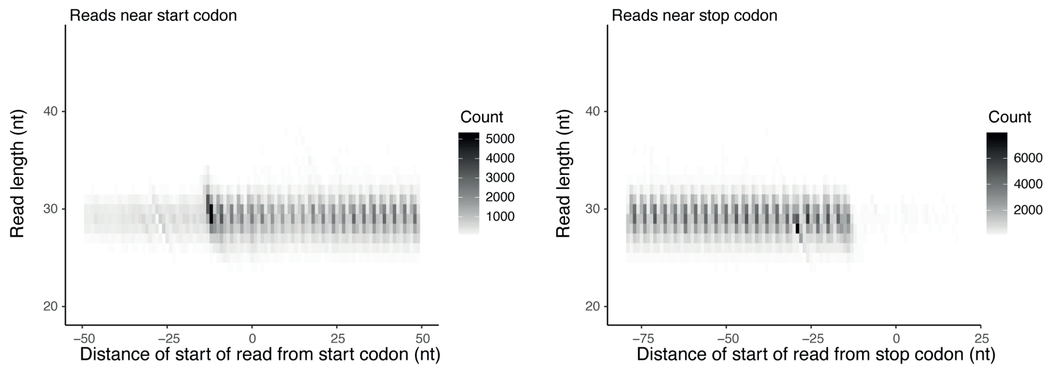


Fig. 6 Heatmaps showing the count and position of 5’ ends of reads around the start (left) and stop (right) codons of transcripts. For ribo-seq, it is expected to see the reduction of signal after the stop codon, an accumulation of read starts upstream of the start codon, and a repeating pattern every 3 nt.

P-site identification and diagnostics

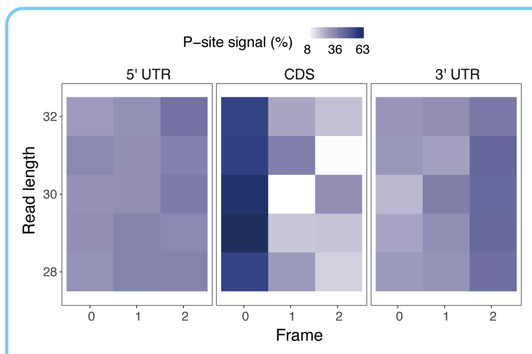


Fig. 7 Heatmaps of P-site percentage in the three frames across untranslated and coding regions generated by riboWaltz[5].

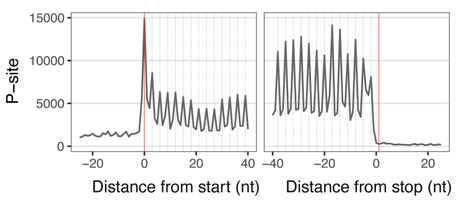


Fig. 8 Meta-profiles showing the periodicity of ribosomes around start and stop codons generated by riboWaltz[5].

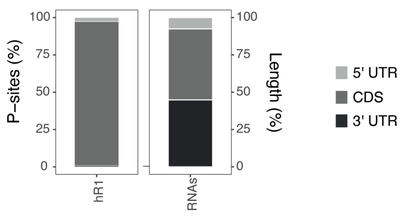


Fig. 9 P-site percentage across protein-coding transcript regions, compared to a theoretical distribution derived from feature lengths (RNAs).

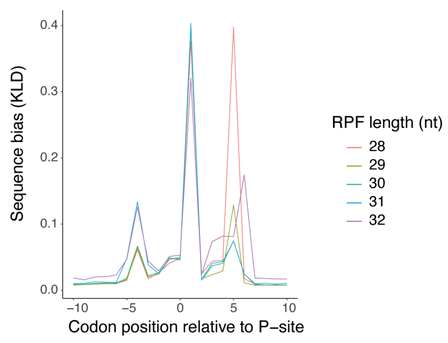


Fig. 10 Ribo-seq Unit Step Transformation (RUST)[6] analysis showing read-length resolved meta-profiles of Kullback–Leibler divergence as a measure of sequence bias.

Interactive visualization for tracking reads throughout the pipeline

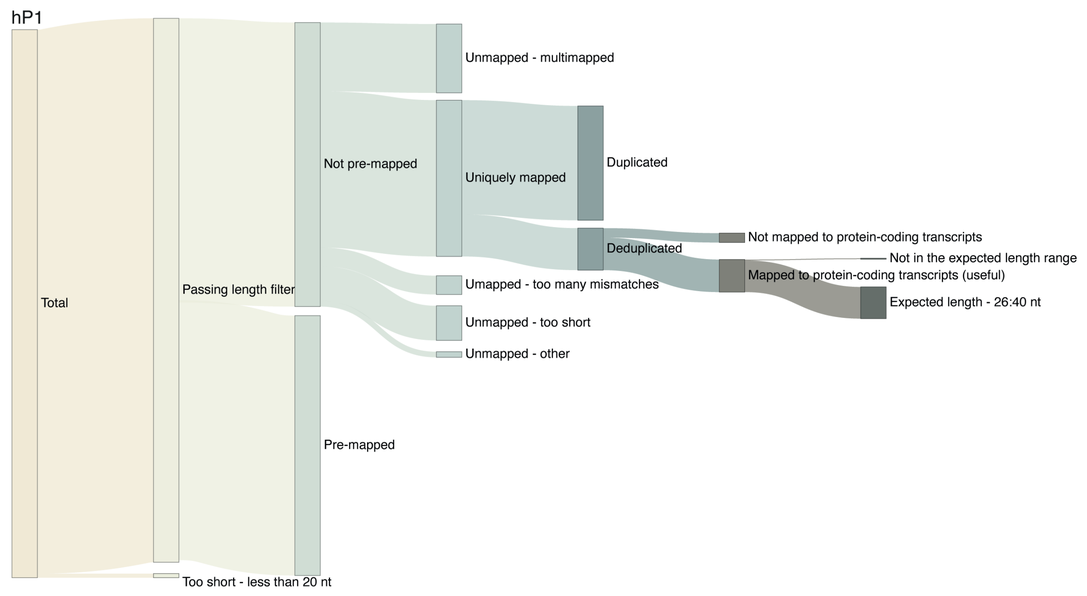


Fig. 11 Sankey diagram showing the read count at each main step of riboseq-flow for one sample.

MultiQC[7] summary reporting for multi-sample Ribo-seq QC

A full example report is available here.

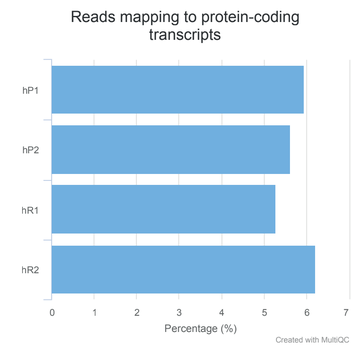


Fig. 12 Percentage of the input reads mapping uniquely to protein-coding transcripts (useful reads).

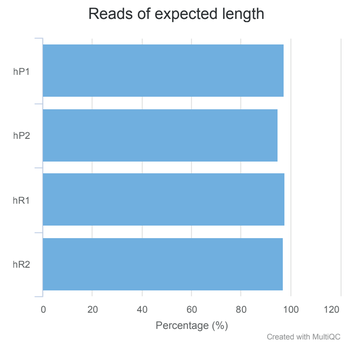


Fig. 13 Percentage of useful reads of expected length. The expected read length range (default: 26-32 nt) can be specified with the `--expected_length` parameter.

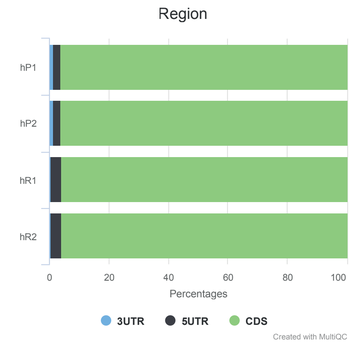


Fig. 14 Distribution of 5' ends of uniquely mapped reads across UTR and CDS regions of protein-coding transcripts.

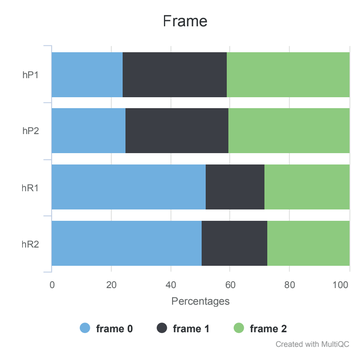


Fig. 14 Frame distribution for the 5' ends of reads in the expected length range.

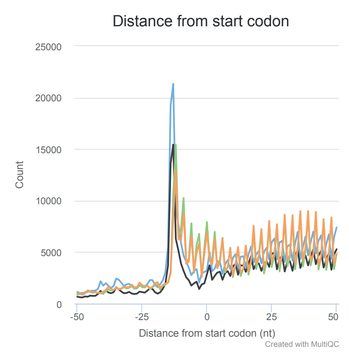


Fig. 15 Metaprofile of 5'ends of useful reads of expected length around start codons.

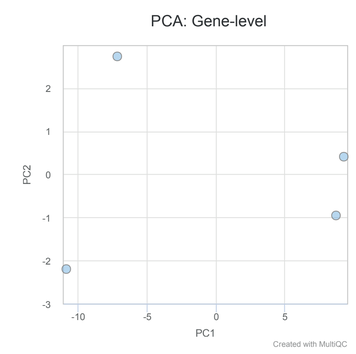


Fig. 16 PCA plot using gene-level counts from uniquely mapped reads.

Effortless and reproducible Ribo-seq analysis with riboseq-flow

- Simplifies adoption of high standards in Ribo-seq analysis, ensuring reproducibility across different computing environments.
- Accessible to researchers without bioinformatics expertise, while providing advanced parameters for experienced users.
- Detailed reporting: per-sample QC reports and MultiQC summary, highlighting key metrics in an interpretable format.
- Transparent read processing: tracks and tallies reads from pipeline logs for clear insight into the data.
- Enables informed analysis: empowers users to make better decisions for downstream analyses or optimisations.
- Results from worked examples with full-sized datasets are available: DOI 10.5281/zenodo.10572576 DOI 10.5281/zenodo.10573243
- Integrative assets: modules and workflow can be integrated into complementary pipelines. Ongoing collaboration with the nf-core community on **nf-core/riboseq**!

REFERENCES

1. Di Tommaso P, Chatzou M, Floden EW, et al.: Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017; 35(4): 316-319.
2. Ewels PA, Peltzer A, Fillinger S, et al.: The nf-core framework for community-curated bioinformatics pipelines. Nat Biotechnol. 2020; 38(3): 276-278.
3. Iosub IA, Wilkins OG, Ule J: Riboseq-flow: A streamlined, reliable pipeline for ribosome profiling data analysis and quality control. Wellcome Open Research. 2024; 9:179
4. Ferguson L, Upton HE, Pimentel SC, et al.: Streamlined and sensitive mono- and di-ribosome profiling in yeast and human cells. Nat Methods. 2023; 20(11): 1704-1715
5. Lauria F, Tebaldi T, Bernabò P, et al.: riboWaltz: Optimization of ribosome P-site positioning in ribosome profiling data. PLoS Comput Biol. 2018; 14(8): e1006169.
6. O'Connor PBF, Andreev DE, Baranov PV: Comparative survey of the relative impact of mRNA features on local ribosome profiling read density. Nat Commun. 2016; 7: 12915
7. Ewels P, Magnusson M, Lundin S, et al.: MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016; 32(19): 3047-3048