Differential analysis of translation efficiency and usage of open reading frames using DOTSeq

Chun Shen Lim[1,2,3,*] and Gabrielle S.W. Chieng[1,2]

[1]Department of Biochemistry, Faculty of Biomedical Sciences, University of Otago, Dunedin, New Zealand
[2]Genetics Otago, University of Otago, Dunedin, New Zealand
[3]Maurice Wilkins Centre for Molecular Biodiscovery, Centres of Research Excellence, New Zealand
[*]Corresponding author. Email: chunshen.lim@otago.ac.nz

Abstract

Protein synthesis is a key cellular process in which mRNAs are translated into proteins by ribosomes. This process is tightly regulated, enabling cells to control protein output in response to specific cellular states. Ribosome profiling captures translatomic landscapes across conditions, but existing computational tools for differential translation analysis operate at the gene level, overlooking translational control at the level of multiple open reading frames (ORFs). Here, we present DOTSeq, a Differential ORF Translation statistical framework that enables systematic discovery of translational control events within genes. DOTSeq offers differential analyses of ORF usage and translation efficiency across biological conditions. These modules allow global detection of *cis*-regulatory events, such as upstream ORF (uORF)-mediated translational control. Benchmarking on simulated datasets demonstrates DOTSeq's sensitivity to subtle regulatory signals, particularly in modest effect sizes where most biological signals occur and where existing tools often show limited sensitivity. DOTSeq provides a flexible and powerful approach for dissecting the complexity of translational control.

Keywords: *cis*-regulatory control, beta-binomial generalised linear model, differential ORF usage, differential translation efficiency, Ribo-seq

1 INTRODUCTION

Protein synthesis is a key step in the gene expression pathway, where mRNAs are translated into functional proteins. The levels of proteins are tightly regulated during translation, a dynamic process that allows cells to adapt to environmental cues and signaling events. With the advent of ribosome profiling (Ribo-seq), we now have the ability to study translation at nucleotide resolution [1, 2]. Ribo-seq is a deep-sequencing

technique that captures ribosome-protected mRNA fragments, typically 30 nucleotides long, following nuclease digestion. These fragments, or footprints, provide a snapshot of ribosome positions along a stretch of nucleotides, thus providing us an estimate of enabling global quantification of translation and identification of reading frames translated.

An open reading frame (ORF) is a continuous stretch of nucleotide sequence that has a length divisible by three and is bounded by start and stop codons [3]. An ORF does not necessarily contain a full coding sequence (CDS) which is a sequence that is eventually translated into a protein. While main ORFs (mORFs) encode canonical protein-coding sequences, small ORFs (300 nt or shorter) have recently gained attention due to widespread ribosome occupancy revealed by Ribo-seq [4, 5, 6]. These small ORF include upstream and downstream ORFs (uORFs and dORFs, respectively) [7]. Some small ORFs encode functional peptides that act as membrane-associated regulators with important cellular roles [8, 6]. Others function as *cis*-regulatory elements, regulating mORF translation through mechanisms such as ribosome stalling, attenuation, or interaction with regulatory proteins. A well-characterised example is ATF4, a key effector of the integrated stress response, where mORF translation is regulated by uORFs [9, 10, 11, 12]. These regulatory mechanisms highlight the complexity of translational control at the level of multiple ORFs. Such complexity, in turn, emphasises the need for genome-wide computational approaches to detect translational regulatory events across biological conditions.

While substantial progress has been made in identifying novel ORFs, such as through our tool `RIBOSS` [13], it is equally important to determine if these ORFs are involved in translational control. However, existing tools for differential translation analysis, including `deltaTE`, `RiboDiff`, `Riborex`, and `Xtail`, operate at the gene level, and cannot resolve differential usage between ORFs within the same gene [14, 15, 16, 17]. To address this limitation, we developed `DOTSeq`, an R package that performs Differential ORF Translation analysis that integrates Ribo-seq and matched RNA-seq data.

`DOTSeq` allows users to ask two key biological questions. (i) Does the translation efficiency of a specific ORF change across conditions? (ii) More importantly, does the relative translation of one ORF compared to others within the same gene change across conditions? `DOTSeq` addresses these two aspects of translational control by modelling both Differential Translation Efficiency (DTE) and Differential ORF Usage (DOU) at the level of multiple ORFs, enabling the detection of regulatory events in response to distinct cellular states. The `DOTSeq` framework provides a powerful approach to dissect the complexity of translational control at the global scale.

2 RESULTS

2.1 Overview of `DOTSeq` modules

Ribo-seq data contains rich information about translation, including complex *cis*-regulatory events such as alternative initiation. These events often contribute to changes in relative usage of ORFs within the same gene, which are not captured by standard gene-level dif-

ferential translation analyses. To address this, we developed DOTSeq, a computational framework that enables global detection of *cis*-translational control events. DOTSeq utilises ORF-level annotations and statistical modelling strategies that capture both absolute changes in translation and condition-specific shifts in the expected proportion of Ribo-seq to RNA-seq read counts for individual ORFs, relative to other ORFs within the same gene.

To enable ORF-level analysis, DOTSeq requires preprocessing steps that include: (i) alignment of Ribo-seq and RNA-seq reads using a splice-aware short-read aligner (e.g., STAR or HISAT2) [18]; and (ii) generation of custom GTF and BED files by "flattening" ORFs across transcript isoforms into discrete, non-overlapping ORF bins using the RIBOSS engine [19, 20] (Figure 1). This step is similar to DEXSeq's step in defining exon counting bins by flattening GTF files [18]. Flattened annotations based on the GENCODE human reference gene annotation [21] are available as part of the package, and custom annotations for other species can be generated using the included orf_to_gtf.py script. Ribo-seq and RNA-seq counts are aggregated per ORF based on the flattened annotation using featureCounts [22].

DOTSeq comprises two statistical modules to model differential usage and translation of ORFs (Table 1). The Differential ORF Usage (DOU) module captures changes in relative usage of ORFs within a gene to identify *cis*-regulatory events. It uses a beta-binomial generalised linear model (GLM) implemented via the glmmTMB package [23]. Specifically, DOU models changes in the expected proportion of Ribo-seq and RNA-seq counts for a given ORF relative to the total counts from other ORFs in the same gene, and tests whether this proportion changes between conditions using an interaction term (condition:strategy). This term captures whether the effect of sequencing strategy (Ribo-seq vs RNA-seq) differs between biological conditions, enabling detection of condition-specific shifts in ORF usage driven by *cis*-translation control. For example, if an ORF's usage remains stable across conditions, its expected proportion of Ribo-seq and RNA-seq counts relative to other ORFs will be consistent, and no condition-specific change in its relative contribution to the gene's overall expression and translation will be detected. However, if an ORF becomes more or less translated in one condition, its expected Ribo-seq proportion will shift relative to its expected RNA-seq proportion and to other ORFs in the gene, indicating a change in ORF usage within the gene context. For this module, dispersion is modelled separately for Ribo-seq and RNA-seq to account for distinct sources of variability. Post hoc analysis includes pairwise contrasts across conditions, computation of estimated marginal means (EMMs) using the emmeans package [24], and adaptive shrinkage of effect sizes via the ashr package [24].

In contrast, the Differential Translation Efficiency (DTE) module quantifies changes in ribosome loading relative to RNA expression at the level of individual ORFs. It is designed to detect DTE in single-ORF/monocistronic genes and in ORFs showing monotonic changes within multi-cistronic genes that the DOU module is not designed to capture. The DTE module uses a similar workflow as deltaTE [17]. This module models raw ORF-level counts directly using a negative binomial generalised linear model (GLM) implemented in DESeq2 [25], and tests whether the ratio of Ribo-seq to RNA-seq counts differs across conditions using an interaction term (condition:strategy). Together,

these complementary modules offer a genome-wide approach for extracting regulatory information in Ribo-seq experiments across biological conditions, enabling systematic interpretation of complex *cis*-regulatory events that are inaccessible to current gene-level tools.

2.2 `DOTSeq` `DOU` and `DTE` modules reveal complementary insights into differential ORF translation

To demonstrate the utility of `DOTSeq`, we analysed publicly available Ribo-seq and matched RNA-seq data for HeLa cells synchronised at two distinct cell cycle stages: mitotic cycling (cells undergoing DNA replication and mitosis) and interphase (cells replicating DNA without mitosis) [26]. The DOU analysis revealed changes in ORF usage within genes that would be missed by a DTE approach alone. In particular, we observed a significant shift towards increased uORF usage in mitotic cycling versus interphase 2A). Such differences highlight the impact of *cis*-regulation, as uORF translation often represses mORFs.

These significant changes include a marked increased uORF usage in the regulatory-associated protein of mTOR (raptor) RPTOR, the initiation factor EIF4G1, and ribosomal protein genes RPL14, and RPS24 (Figure 2A). This pattern is consistent with previous studies, showing that global translation is repressed during mitosis, at least partly to a greater extent than transcription [27, 28, 29, 30]. While this repression has been attributed to raptor and eIF4G1 phosphorylation [28, 31, 32, 33], our findings suggest an additional layer of translational control mediated by uORFs. In contrast to mitotic cycling, RPTOR and EIF4G1 mORF usage is significantly elevated during interphase (Figure 2A). This observation aligns with our understanding that raptor and eIF4G1 are two master regulators of protein synthesis [34, 35]. Raptor is part of the mTOC1 complex that promotes protein synthesis, cell growth and metabolism [34, 36], whereas eIF4G1 is part of the cap-binding complex that mediates recruitment of ribosomes to mRNAs [37]. Together, these findings suggest that the RPTOR and EIF4G1 are regulated through multiple layers, with preferential mORF usage during interphase, supporting its role in cell-cycle progression.

In contrast to these uORF-mediated repression events, we observed a shift toward mORF usage in the H2AC6 gene, an isotype of the histone 2A, during mitotic cycling. Previous studies have shown that H2A deubiquitination is required for proper progression through the M phase [38]. Thus, the shift in mORF usage supports the notion that H2AC6 plays a crucial role in mitotic cycling. Taken together, these results demonstrate that the `DOU` module is capable of revealing *cis*-translational control events.

To assess whether the `DOU` and `DTE` modules capture overlapping or distinct aspects of translational control, we compared effect sizes between the two modules using the HeLa cell cycle data above. Most changes in translation efficiency fell within a modest range (± 1 log2-fold change), where 153 of differentially translated ORFs were missed by the `DTE` module but detected by `DOU` (Figure 2B and 2C). However, there was a strong correlation between DOU log-odds changes and DTE log-fold changes, indicating that both modules detect condition-dependent changes in translation through different modelling approaches (Figure 2B, Spearman $\rho = 0.65$, p-value $< 10^{-16}$). Notably, the

overlap between the ORFs meeting statistical thresholds was partial: 309 ORFs were unique to DTE, 153 unique to DOU, and 243 shared (Figure 2C, overlap coefficient = 0.56). These findings demonstrate that the two modules provide complementary insights rather than redundant results.

2.3 `DOU` and `DTE` modules show complementary strengths in simulation benchmarks

To evaluate performance under controlled conditions, we designed simulation scenarios modelling DOU using the `DOTSeq` `simDOT` function, which builds on the `polyester` simulation engine [39]. These scenarios incorporated experimentally derived Ribo-seq and matched RNA-seq count matrices from the HeLa cell cycle dataset above [26]. To assess sensitivity across biological relevant effect sizes (Figure 2B), we introduced varying magnitudes of log-fold changes in translation efficiency to multi-cistronic genes. These include simulating the absolute magnitude differences between uORFs and mORFs from 1 to 3 log-fold changes, representing low to large effect sizes, respectively (Figure 3, 4, and Supplementary Figure S1-4). We also incorporated batch effects to provide realistic experimental variability.

We analysed each simulation using the `DOU` and `DTE` modules, and benchmarked them against existing tools for differential translation analysis, including `deltaTE` [17], `RiboDiff` [15], `Riborex` [17], and `Xtail` [14]. Performance was assessed using the `iCOBRA` package, including False Discovery Rate (FDR) and True Positive Rate (TPR) across multiple thresholds [40].

The `DOU` module outperformed the `DTE` module and other tools at low and moderate effect sizes where most biological signals occur (Figure 3 and 4). Although the `DOU` module slightly exceeds the nominal FDR thresholds in all cases, it consistently produces the narrowest range of observed FDR values, indicating tight control and conservative behavior. In contrast, other tools show greater variability, with only several results falling below the nominal threshold at 0.1 and 0.05. This demonstrates the power of the modelling of condition-specific changes in the expected proportional counts for an ORF relative to other ORFs within the same gene via a beta-binomial framework.

While existing tools exhibited limited sensitivity at low effect sizes where the signal-to-noise ratio is low, performance improved with increasing effect size. The `DTE` module and other tools showed greater sensitivity to effect size and outperformed the `DOU` module at the large magnitudes (absolute log-fold change >2 between uORFs and mORFs within genes; Supplementary Figure S1, S3, and S4). The `DTE` module is functionally equivalent to `deltaTE` and, as expected, exhibited identical performance in our benchmarks. Together, these results highlight the combined utility of the two `DOTSeq` modules in capturing a broad spectrum of translational control events.

To further assess statistical calibration, we examined the distribution of significance measures produced by the `DOTSeq` modules across 11 scenarios modelling differential ORF usage. These scenarios varied in absolute effect size and included two regulatory patterns: uORF up with mORF down, and uORF down with mORF up. The `DOU` module produced a right-skewed p-value distribution, indicating well-calibrated statistical significance (Figure 5). In contrast, the `DTE` module exhibited a U-shaped distribution,

suggesting p-value inflation and reduced calibration when applied to DOU-type regulatory events. These results highlight the importance of using DOU-specific models for detecting translational control events involving DOU.

3 DISCUSSION

The ability to resolve changes in ORF usage across biological conditions opens up new avenues to exploring translational control mechanisms that are obscured in gene-level analyses. `DOTSeq` addresses this need by enabling differential analyses of both ORF usage and translation efficiency. This enhanced sensitivity and resolution enables discovery of important regulatory events (e.g., uORF-mediated repression) that are often overlooked in gene-level analyses.

For example, we discovered key *cis*-regulatory events contributing to global translation repression during mitosis, including uORF-mediated repression of RPTOR and EIF4G1, two key master regulators of protein synthesis, and several ribosomal protein genes. These findings highlight `DOTSeq`'s ability to uncover an additional regulatory layer of protein synthesis in Ribo-seq experiments.

To evaluate `DOTSeq`'s ability in capturing these regulatory events, we have conducted simulations to benchmark its performance. These simulations demonstrate the complementary strengths of the two `DOTSeq` modules: the `DOU` module is effective at detecting subtle, condition-specific shifts in ORF usage in noisy datasets, whereas the `DTE` module is more sensitive to large-magnitude changes in translation efficiency. While low sensitivity at modest effect sizes remains a common challenge, the `DOU` module consistently outperforms existing tools in this regime, even with high technical noise.

These discoveries rely on `DOTSeq`'s ability to statistically model overdispersed and sparse count data from Ribo-seq and RNA-seq using the state-of-the-art TMB (Template Model Builder) architecture [23]. However, fitting beta-binomial GLMs on sparse Ribo-seq data can lead to convergence issues, including a non-positive definite Hessian. To mitigate these issues, `DOTSeq` supports a brute force optimisation step using three optimisers implemented in `glmmTMB`: `nlminb`, `bobyqa`, and `optim`. In addition, we have implemented multiple dispersion modelling approaches within our framework. (i) The default approach models dispersion as a function of sequencing strategy (Ribo-seq with RNA-seq as reference), allowing for strategy-dependent overdispersion. Although this approach performs well in simulations, it can be more prone to convergence issues than simpler dispersion models. (ii) As an alternative, users may assume dispersion to be constant across all predictor levels. This approach reduces model complexity and is less prone to convergence issues. (iii) To balance model stability and flexibility, both models are fitted and compared using likelihood ratio tests. The simpler model is selected when no significant difference is detected. (iv) For maximum flexibility, users may specify a custom formula for dispersion modelling. Random effects can be included in the mean formula and dispersion formulas within the framework generalised linear mixed modelling (GLMM). Users may also enable model diagnostics to assess model fit, including tests for overdispersion, zero inflation, and residual properties.

While our framework is flexible, fitting fully parameterised models comes with a

computational cost. For example, it took 6 min 53 s (± 42 s) to fit the beta-binomial GLMs for 35,863 ORFs in the HeLa cell cycle dataset, running on a single thread of a MacBook Pro M4 with 16 GB RAM. In contrast, the `DTE` module runs at a speed comparable to existing tools such as `deltaTE` (31 ± 2 s). To mitigate this computational cost, `DOTSeq` supports parallel processing via `OpenMP`, which is used internally in `glmmTMB` to accelerate model fitting. Altogether, `DOTSeq` offers a flexible and powerful framework for dissecting the complex landscape of translational control. `DOTSeq` is broadly applicable to studies of translational regulation across conditions within biological contexts such as development, disease, and stress responses.

4 METHODS

4.1 Data

The Ribo-seq and matched RNA-seq data for HeLa cells synchronised at interphase, mitotic arrest, and mitotic cycling were retrieved from PRJNA957808 [26]. There are two biological replicates for each condition.

4.2 Data Preprocessing

Adapter trimming was performed using `Cutadapt` v5.1 [41]. Reads were mapped to the UCSC Genome Browser's hg38 with GENCODE v47 annotation using `STAR` v2.7.11b [19, 42, 21].

For ORF-level quantification, flattened GTF and BED files containing ORF annotations were generated by processing the GENCODE annotation and transcript FASTA files using the `orf_to_gtf.py` script from `DOTSeq`, with the `RIBOSS` v0.45 engine [13]. Reads mapped to individual ORFs were counted using `featureCounts` v2.1.1 [22], guided by this flatten GTF file. The resulting flatten annotation files and count matrices were used as input for `DOTSeq`.

4.3 Statistical modelling for DOU

The `DOTSeq` framework builds on a beta-binomial GLM implemented in `glmmTMB` v1.1.12 [23]. The expected proportion of reads mapped to an ORF within its gene is modeled by a mean model and a dispersion model. The mean model uses the logit link function to model the expected probability (μ_i) that a read maps to a particular ORF for observation i.

4.3.1 Testing the interaction effect

`DOTSeq` provides an optional test to assess the significance of a translation-specific effect. A likelihood ratio test can be performed by comparing a full model, which includes an interaction between condition and strategy, against a null model, which contains only

additive effects:

$$\text{Model}_{\text{full}}: \quad \text{logit}(\mu_i) = \beta_0 + \beta_{\text{cond}}\, x_{\text{cond},i} + \beta_{\text{strat}}\, x_{\text{strat},i} + \beta_{\text{int}}\, (x_{\text{cond},i} \cdot x_{\text{strat},i}) \tag{1}$$

$$\text{Model}_{\text{null}}: \quad \text{logit}(\mu_i) = \beta_0 + \beta_{\text{cond}}\, x_{\text{cond},i} + \beta_{\text{strat}}\, x_{\text{strat},i} \tag{2}$$

The test compares the log-likelihoods of the full and null models to assess whether the inclusion of the interaction term, β_{int}, significantly improves model fit.

The unified mean model (Equation 1) can be simplified by substituting the dummy variable values to isolate the effective intercept and slope for each sequencing level. For the RNA-seq level ($x_{\text{strat},i} = 0$), the model simplifies to:

$$\text{logit}(\mu_{\text{rna},i}) = \beta_0 + \beta_{\text{cond}}\, x_{\text{cond},i} \tag{3}$$

For the Ribo-seq level ($x_{\text{strat},i} = 1$), the model simplifies to:

$$\text{logit}(\mu_{\text{ribo},i}) = (\beta_0 + \beta_{\text{strat}}) + (\beta_{\text{cond}} + \beta_{\text{int}})\, x_{\text{cond},i} \tag{4}$$

The term β_{int} captures the differential effect of condition between the Ribo-seq and RNA-seq levels. Therefore, when $\beta_{\text{int}} \neq 0$, the condition affects Ribo-seq differently than RNA-seq, indicating a translation-specific effect (DOU).

4.3.2 Contrasts

Since `DOTSeq` models the expected proportion of reads, a condition-dependent contrast between Ribo-seq and RNA-seq levels is performed using `emmeans` v1.11.2-8 with Wald test [43] to isolate these translation-specific effects:

$$\text{logit}(\mu_{\text{ribo},i}) - \text{logit}(\mu_{\text{rna},i}) \;=\; \beta_{\text{strat}} + \beta_{\text{int}}\, x_{\text{cond},i} \tag{5}$$

This contrast quantifies how much more (or less) likely a read is to map to an ORF at the Ribo-seq level compared to the RNA-seq level, condition-dependently. While not equivalent to DTE, DOU reflects a condition-dependent changes in ORF usage across sequencing strategies, capturing *cis*-regulatory effects on translation.

For statistical inference, Wald tests are used for condition-specific contrasts, and adaptive empirical Bayesian shrinkage is applied to the log-odds contrasts using `ashr` v2.2-63 [24] to stabilise effect size estimates.

4.3.3 Dispersion modelling

To account for overdispersion in proportional modelling, `DOTSeq` models variability using the dispersion model in a log link function:

$$\log(\phi_i) \;=\; \gamma_{0,i} + \gamma_{1,i}\, x_{\text{strat},i} \tag{6}$$

where ϕ_i is the overdispersion parameter for ORF_i, γ_0 is the dispersion intercept, and γ_1 captures the additional dispersion at the Ribo-seq level. This allows the model to fit

a shared mean structure across Ribo-seq and RNA-seq levels and estimate separate dispersion parameters for these two levels. The fitted dispersions for ORF_i are computed as:

$$\phi_{\text{rna},i} = \exp(\gamma_{0,i}), \quad \phi_{\text{ribo},i} = \exp(\gamma_{0,i} + \gamma_{1,i}) \tag{7}$$

where $\gamma_{0,i}$ and $\gamma_{1,i}$ are the intercept and strategy effect for ORF_i, respectively. This approach accommodates varying technical noise and biological heterogeneity across the sequencing levels, thereby improving statistical inference accuracy.

4.4 Statistical modelling for DTE

The `DTE` module in `DOTSeq` uses the `DESeq2` framework to model DTE across conditions directly through the `condition:strategy` interaction term [25, 17]. For each ORF, raw Ribo-seq and RNA-seq read counts are normalised and modelled using a negative binomial GLM. This includes a standard `DESeq2` workflow: size factor estimation, dispersion fitting, Wald tests for condition-specific contrasts, and shrinkage of effect sizes. This complements the `DOU` module by identifying ORFs with overall changes in translation, using transcript abundance as baseline.

4.5 Simulation

To simulate *cis*-translational control, 11 datasets were generated using the `DOTSeq`'s simDOT function, which is based on the `deltaTE` source code [17] and uses the `polyester` v1.29.1 engine [39]. The HeLa cell cycle count matrices were used as input to preserve realistic baseline expression profiles.

Simulation scenarios included predefined regulatory events, namely "uORF_up_mORF_down" and "uORF_down_mORF_up", which assigned log-fold changes ranging from ± 0.5 to ± 1.5 to specific ORF types within multi-cistronic genes (absolute magnitude differences from 1 to 3 log-fold changes). To simulate experimental variability, a batch effect was introduced to 10% of ORFs, with random log-fold changes centered around 0.9 applied to the non-reference batches. Each simulation included two experimental conditions, with 2 biological replicates per condition and 2 batches. The simulation output comprised combined count matrices for Ribo-seq and RNA-seq, sample-level metadata (including condition, replicate, strategy, and batch), ground truth labels for differentially regulated ORFs, and the corresponding log-fold changes.

4.6 Benchmarking and analysis

Benchmarks were performed on a Ubuntu 24.04.2 LTS server equipped with 2 $\times$ AMD EPYC 7763 64-Core Processors (128 physical cores, 128 threads), running at up to 3.53 GHz, with 1 TB DDR4 RAM. The performance of `DOTSeq` was benchmarked using simulation datasets against four existing tools for detecting differential translation: `deltaTE` [17], `RiboDiff` v0.2.2 [15], `Riborex` v2.4.0 [16], and `Xtail` v.1.20 [14]. The `DOU` and `DTE` modules of `DOTSeq` were evaluated independently.

Performance was assessed using `iCOBRA` v1.37.1 [40], which was used to generate FDR-TPR curves. The `onlyshared` argument in the `calculate_performance` function was set to FALSE as default, allowing all features in the truth table to be used, irrespective of whether they were detected by all methods.

Overlap between significant ORFs identified by the `DOU` and `DTE` modules was visualised using `eulerr` v7.0.2 [44]. Additional plots were produced using the base R functions [45] and `ggplot2` v3.5.2 [40].

4.7 Use of LLMs

LLMs were used to generate docstrings, improve clarity, and/or identify inconsistencies in the code and the main text.

4.8 Code availability

`DOTSeq` is available at https://github.com/compgenom/DOTSeq.git.

5 AUTHOR CONTRIBUTIONS

C.S.L. conceptualised and led the study, developed the statistical framework and conducted data analysis. G.S.W.C. performed benchmarking analyses, prepared the GitHub repository, and developed the package vignette. C.S.L. and G.S.W.C. wrote the manuscript.

6 ACKNOWLEDGEMENTS

The authors appreciate Assoc Prof Chris M. Brown's feedback on the manuscript.

7 Conflicts of interest

The authors declare that they have no conflicts of interest.

8 FUNDING

This work was supported by a Marsden Fund Fast-Start Grant (MFP-UOO-2111) from Government funding, administered by the Royal Society of New Zealand Te Apārangi, and by research funds from a University of Otago Research Grant and the Otago School of Biomedical Sciences Dean's Fund awarded to C.S.L.

References

[1] Nicholas T Ingolia et al. "Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling". In: *science* 324.5924 (Apr. 2009), pp. 218–223.

[2] Kotaro Tomuro and Shintaro Iwasaki. "Advances in ribosome profiling technologies". In: *Biochemical Society Transactions* (May 2025), BST20253061.

[3] Patricia Sieber, Matthias Platzer, and Stefan Schuster. "The Definition of Open Reading Frame Revisited". en. In: *Trends Genet* 34.3 (Mar. 2018), pp. 167–170.

[4] Jonathan M Mudge et al. "Standardized annotation of translated open reading frames". en. In: *Nat Biotechnol* 40.7 (July 2022), pp. 994–999.

[5] Eric W Deutsch et al. "High-quality peptide evidence for annotating non-canonical open reading frames as human proteins". en. Sept. 2024.

[6] Roger P Hellens et al. "The Emerging World of Small ORFs". en. In: *Trends Plant Sci* 21.4 (Apr. 2016), pp. 317–328.

[7] Qiushuang Wu et al. "Translation of small downstream ORFs enhances translation of canonical main open reading frames". en. In: *EMBO J* 39.17 (Sept. 2020), e104763.

[8] Casimiro Baena-Angulo, Ana Isabel Platero, and Juan Pablo Couso. "Cis to trans: small ORF functions emerging through evolution". en. In: *Trends Genet* 41.2 (Feb. 2025), pp. 119–131.

[9] Krishna M Vattem and Ronald C Wek. "Reinitiation involving upstream ORFs regulates ATF4 mRNA translation in mammalian cells". en. In: *Proc Natl Acad Sci U S A* 101.31 (Aug. 2004), pp. 11269–11274.

[10] Phoebe D Lu, Heather P Harding, and David Ron. "Translation reinitiation at alternative open reading frames regulates gene expression in an integrated stress response". en. In: *J Cell Biol* 167.1 (Oct. 2004), pp. 27–33.

[11] Graham Neill and Glenn R Masson. "A stay of execution: ATF4 regulation and potential outcomes for the integrated stress response". en. In: *Front Mol Neurosci* 16 (Feb. 2023), p. 1112253.

[12] Anna M Smirnova et al. "Stem-loop-induced ribosome queuing in the uORF2/ATF4 overlap fine-tunes stress-induced human ATF4 translational control". en. In: *Cell Rep* 43.4 (Apr. 2024), p. 113976.

[13] Chun Shen Lim et al. "RIBOSS detects novel translational events by combining long- and short-read transcriptome and translatome profiling". en. In: *Brief Bioinform* 26.2 (Mar. 2025).

[14] Zhengtao Xiao et al. "Genome-wide assessment of differential translations with ribosome profiling data". en. In: *Nat Commun* 7 (Apr. 2016), p. 11194.

[15] Yi Zhong et al. "RiboDiff: detecting changes of mRNA translation efficiency from ribosome footprints". en. In: *Bioinformatics* 33.1 (Jan. 2017), pp. 139–141.

[16] Wenzheng Li et al. "Riborex: fast and flexible identification of differential translation from Ribo-seq data". en. In: *Bioinformatics* 33.11 (June 2017), pp. 1735–1737.

[17] Sonia Chothani et al. "deltaTE: Detection of Translationally Regulated Genes by Integrative Analysis of Ribo-seq and RNA-seq Data". en. In: *Curr Protoc Mol Biol* 129.1 (Dec. 2019), e108.

[18] Simon Anders, Alejandro Reyes, and Wolfgang Huber. "Detecting differential usage of exons from RNA-seq data". en. In: *Genome Res* 22.10 (Oct. 2012), pp. 2008–2017.

[19] Alexander Dobin et al. "STAR: ultrafast universal RNA-seq aligner". en. In: *Bioinformatics* 29.1 (Jan. 2013), pp. 15–21.

[20] Daehwan Kim et al. "Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype". en. In: *Nat Biotechnol* 37.8 (Aug. 2019), pp. 907–915.

[21] Jonathan M Mudge et al. "GENCODE 2025: reference gene annotation for human and mouse". en. In: *Nucleic Acids Res* 53.D1 (Jan. 2025), pp. D966–D975.

[22] Yang Liao, Gordon K Smyth, and Wei Shi. "featureCounts: an efficient general purpose program for assigning sequence reads to genomic features". en. In: *Bioinformatics* 30.7 (Apr. 2014), pp. 923–930.

[23] Maeve McGillycuddy et al. "Parsimoniously fitting large multivariate random effects in glmmTMB". en. In: *J. Stat. Softw.* 112.1 (2025).

[24] Matthew Stephens. "False discovery rates: a new deal". en. In: *Biostatistics* 18.2 (Apr. 2017), pp. 275–294.

[25] Michael I Love, Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". en. In: *Genome Biol* 15.12 (2014), p. 550.

[26] Jimmy Ly et al. "Nuclear release of eIF1 restricts start-codon selection during mitosis". en. In: *Nature* 635.8038 (Nov. 2024), pp. 490–498.

[27] Marvin E Tanenbaum et al. "Regulation of mRNA translation during mitosis". en. In: *Elife* 4 (Aug. 2015).

[28] Mikhail I Dobrikov et al. "Mitotic phosphorylation of eukaryotic initiation factor 4G1 (eIF4G1) at Ser1232 by Cdk1:cyclin B inhibits eIF4A helicase complex binding with RNA". en. In: *Mol Cell Biol* 34.3 (Feb. 2014), pp. 439–451.

[29] A Contreras and C Perea-Resa. "Transcriptional repression across mitosis: mechanisms and functions". en. In: *Biochem Soc Trans* 52.1 (Feb. 2024), pp. 455–464.

[30] Qinyu Hao et al. "The S-phase-induced lncRNA SUNO1 promotes cell proliferation by controlling YAP1/Hippo signaling pathway". en. In: *eLife* (Oct. 2020).

[31] Dana M Gwinn et al. "AMPK phosphorylation of raptor mediates a metabolic checkpoint". In: *Molecular cell* 30.2 (2008), pp. 214–226.

[32] Mohamed Moustafa-Kamal et al. "The mTORC1/S6K/PDCD4/eIF4A axis determines outcome of mitotic arrest". In: *Cell Reports* 33.1 (2020).

[33] Richard I Odle et al. "An mTORC1-to-CDK1 switch maintains autophagy suppression during mitosis". In: *Molecular cell* 77.2 (2020), pp. 228–240.

[34] Long He, Sungyun Cho, and John Blenis. "mTORC1, the maestro of cell metabolism and growth". In: *Genes & Development* 39.1-2 (2025), pp. 109–131.

[35] Nathan R James and John S O'Neill. "Circadian Control of Protein Synthesis". In: *BioEssays* 47.3 (2025), e202300158.

[36] Jay N Joshi et al. "mTORC1 activity oscillates throughout the cell cycle, promoting mitotic entry and differentially influencing autophagy induction". In: *Cell Reports* 43.8 (2024).

[37] Jailson Brito Querido, Irene Díaz-López, and V Ramakrishnan. "The molecular basis of translation initiation and its regulation in eukaryotes". In: *Nature Reviews Molecular Cell Biology* 25.3 (2024), pp. 168–186.

[38] Heui-Yun Joo et al. "Regulation of cell cycle progression and gene expression by H2A deubiquitination". en. In: *Nature* 449.7165 (Oct. 2007), pp. 1068–1072.

[39] Alyssa C Frazee et al. "Polyester: simulating RNA-seq datasets with differential transcript expression". en. In: *Bioinformatics* 31.17 (Sept. 2015), pp. 2778–2784.

[40] Charlotte Soneson and Mark D Robinson. "iCOBRA: open, reproducible, standardized and live method benchmarking". en. In: *Nat Methods* 13.4 (Apr. 2016), p. 283.

[41] Marcel Martin. "Cutadapt removes adapter sequences from high-throughput sequencing reads". In: *EMBnet J.* 17.1 (May 2011), p. 10.

[42] Gerardo Perez et al. "The UCSC Genome Browser database: 2025 update". en. In: *Nucleic Acids Res* 53.D1 (Jan. 2025), pp. D1243–D1249.

[43] Russell V. Lenth. *emmeans: Estimated Marginal Means, aka Least-Squares Means.* R package version 1.11.2-80003. 2025. URL: `https://rvlenth.github.io/emmeans/`.

[44] Johan Larsson. "Area-Proportional Euler and Venn Diagrams with Ellipses [R package eulerr version 7.0.2]". In: (Mar. 2024).

[45] *The R Project for Statistical Computing*. en. `https://www.R-project.org/`. Accessed: 2025-9-21.

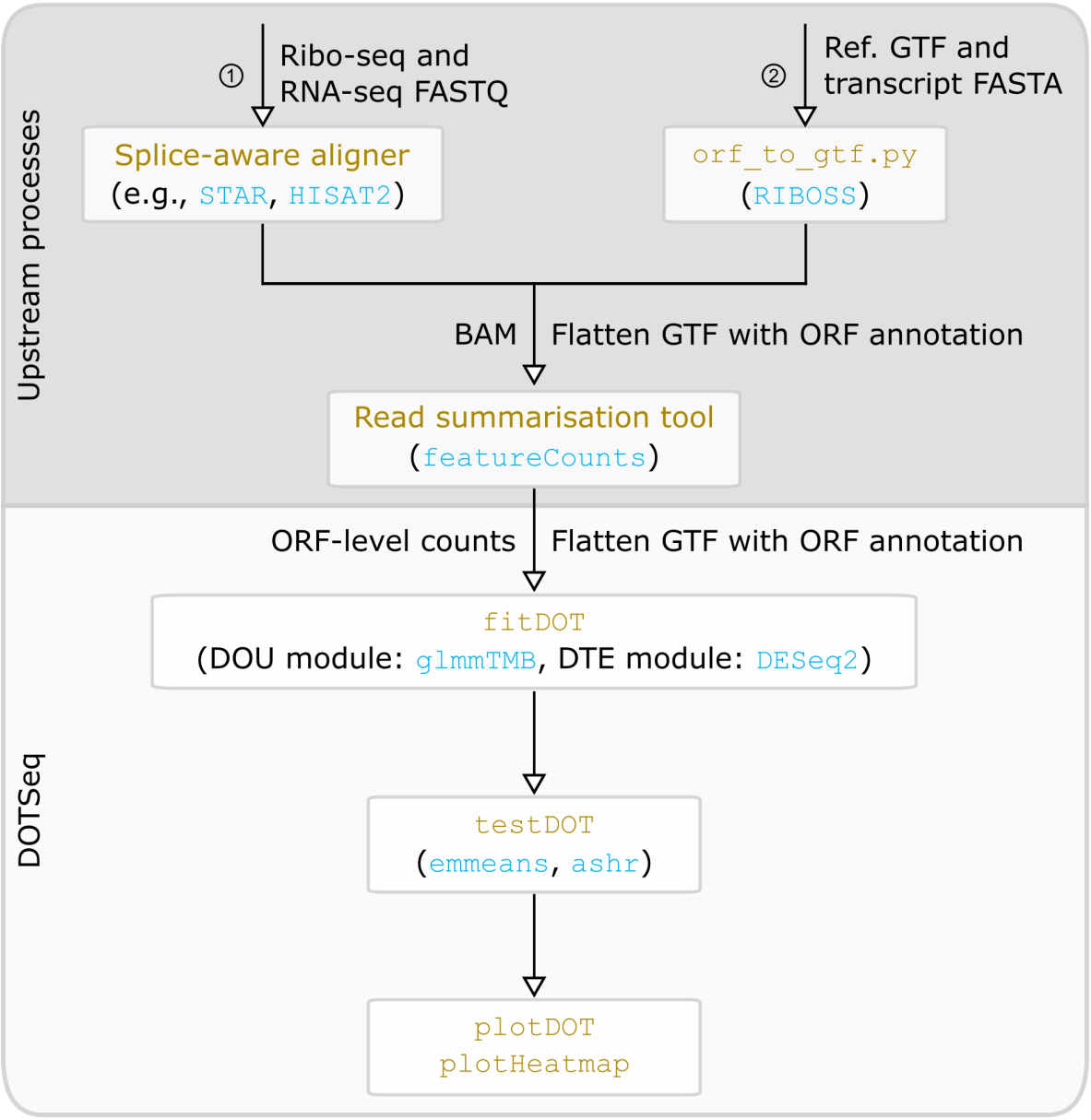

Figure 1: Overview of the `DOTSeq` workflow. ① Ribo-seq and matched RNA-seq reads are aligned to a reference genome using a splice-aware aligner, such as `STAR` or `HISAT2`. ② Reference GTF and transcript FASTA files are parsed using the `orf_to_gtf.py` wrapper, which leverages the `RIBOSS` engine to generate flattened GTF and BED files with discrete, non-overlapping ORF annotations. `DOTSeq` integrates two statistical modules: Differential ORF Usage (`DOU`) and Differential Translation Efficiency (`DTE`). The `fitDOT` function performs model fitting, while `testDOT` enables post hoc contrasts and empirical Bayesian shrinkage. The `plotDOT` function generates a scatter plot of DOU vs DTE effect size estimates and a Venn diagram showing overlap between DOU and DTE results. The `plotHeatmap` function generates a heatmap of genes with DOU.

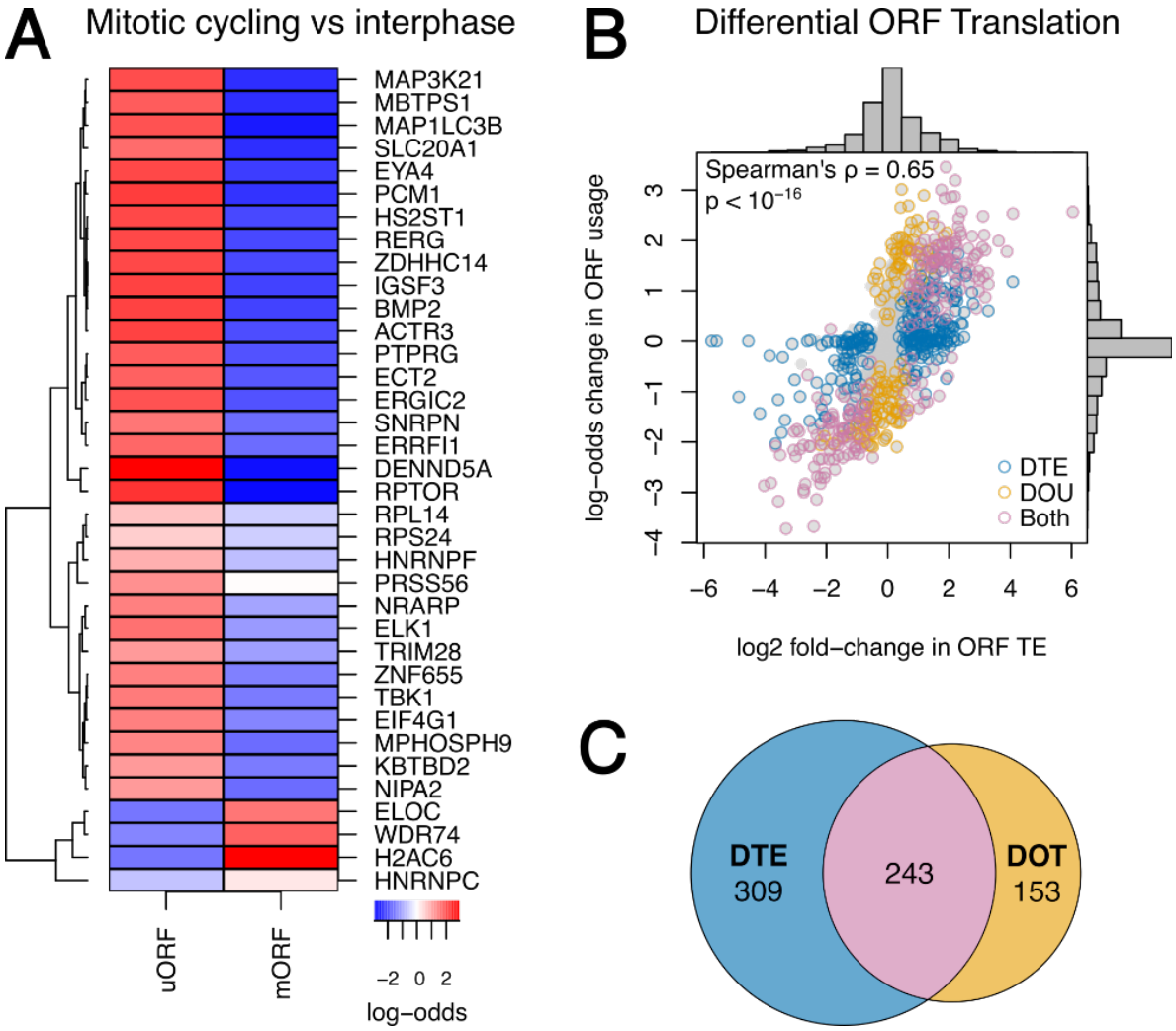

Figure 2: `DOTSeq` reveals distinct patterns of differential ORF translation between mitotic cycling and interphase. Analysis of Ribo-seq and RNA-seq data comparing HeLa cells synchronised at interphase and mitotic cycling. **(A)** Heatmap of genes with DOU, showing log-odds values ranging approximately from −2 to +2. Rows represent genes, highlighting differences in uORF and mORF usage between mitotic cycling and interphase conditions. **(B)** Scatter plot showing a significant positive correlation between DOU log-odds changes and DTE log-fold changes (Spearman $\rho = 0.65$, p-value $< 10^{-16}$). Effect sizes for both DOU and DTE have been shrunk using empirical Bayesian estimators to improve stability and interpretability. **(C)** Venn diagram illustrating overlap between ORFs detected by the `DOU` and `DTE` modules. Of the total, 309 ORFs are unique to DTE, 153 to DOU, and 243 are shared, yielding an overlap coefficient of 0.59. Plots were generated using the `plotHeatmap` and `plotDOT` functions in `DOTSeq`.

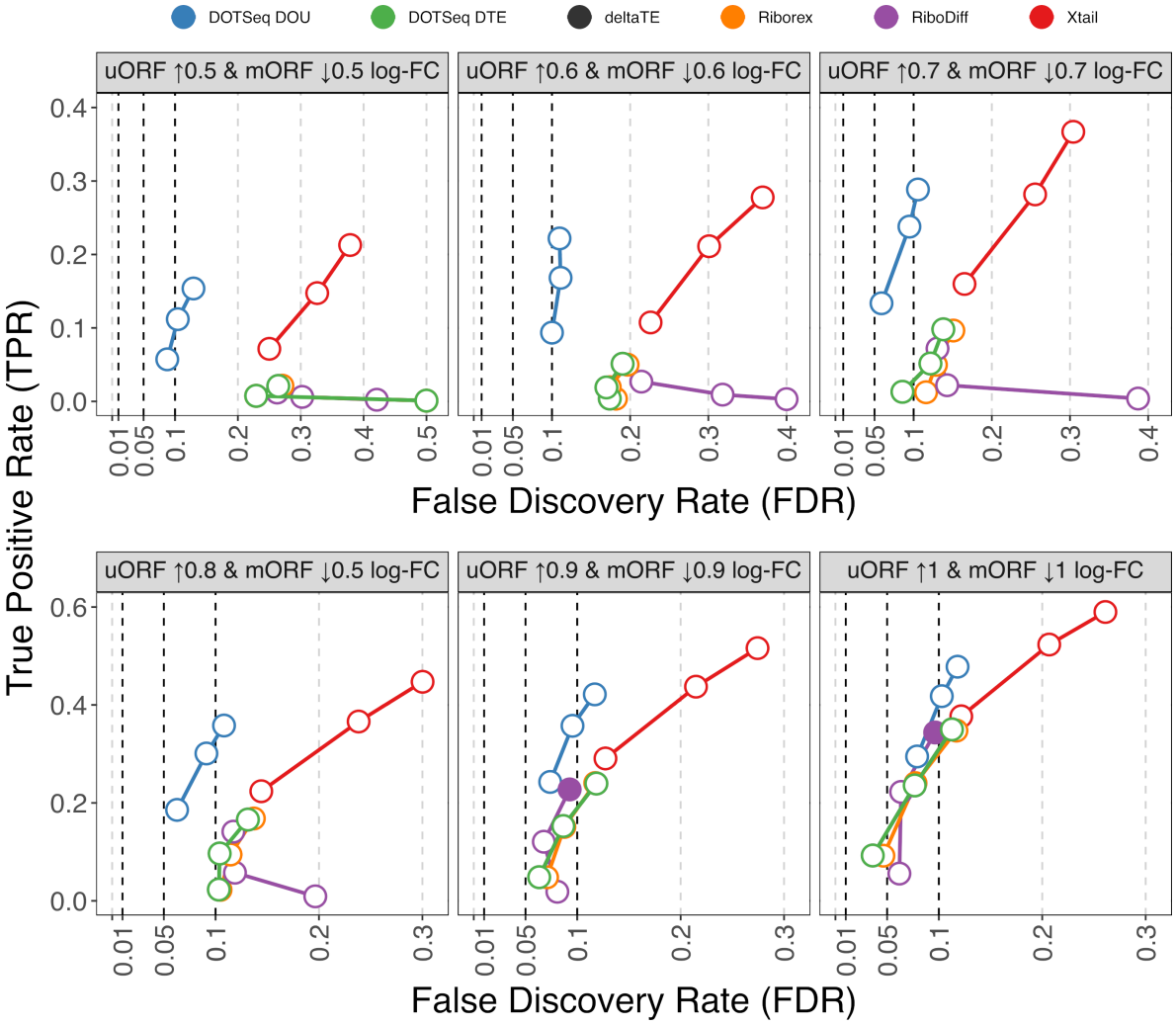

Figure 3: DOTSeq detects simulated *cis*–translational control events at varying effect sizes. False Discovery Rate–True Positive Rate (FDR–TPR) plots compare the performance of DOTSeq's DOU and DTE modules against existing tools across biologically realistic effect sizes (Figure 2), including absolute magnitude differences between uORFs and mORFs ranging from 1 to 2 log-fold change (log-FC). Each simulation included Ribo-seq and matched RNA-seq data across two conditions, each with two batches and two biological replicates (eight samples total). To introduce realistic experimental variability, a batch effect was applied to 10% of ORFs, with random log-fold changes centred around 0.9 applied to the non-reference batches. Each point represents performance at nominal FDRs of 0.01, 0.05, and 0.1, and are filled if the observed FDR is below the threshold. The points for the DOU module are closest to the top-left corner of the plot, indicating a higher TPR at a given observed FDR compared to the other tools.

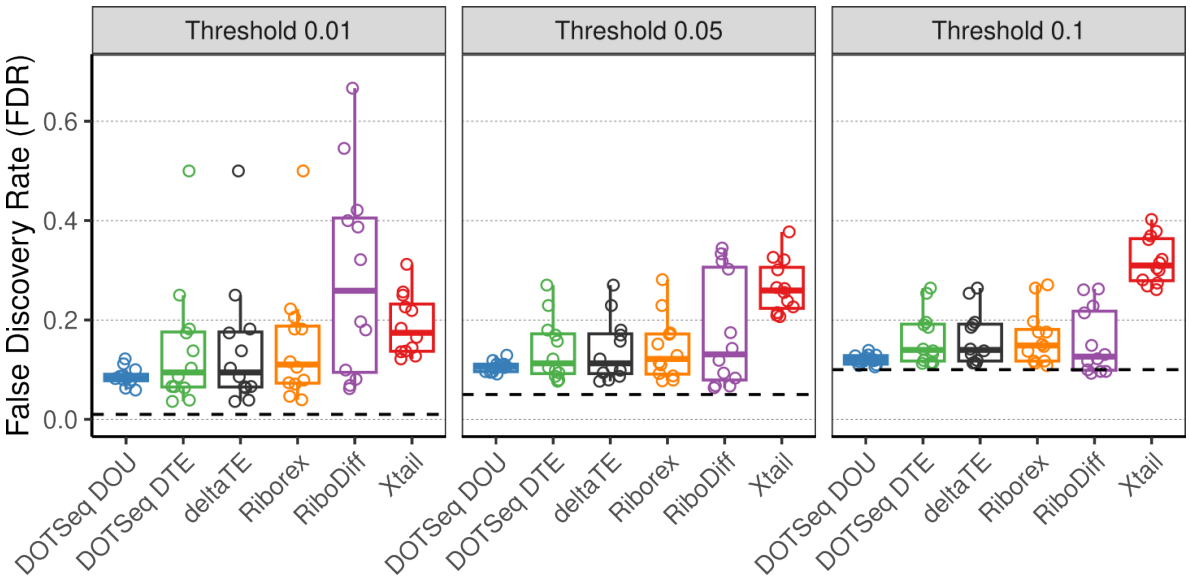

Figure 4: `DOTSeq DOU` module shows the most consistent and conservative control of FDR for differential translation analysis. The performance of the `DOTSeq DOU` and `DTE` modules was benchmarked against existing tools at nominal FDR thresholds of 0.01, 0.05, and 0.1. Each point in the box plots represents the observed FDR at a given nominal threshold, while the y-axis shows the actual FDR.

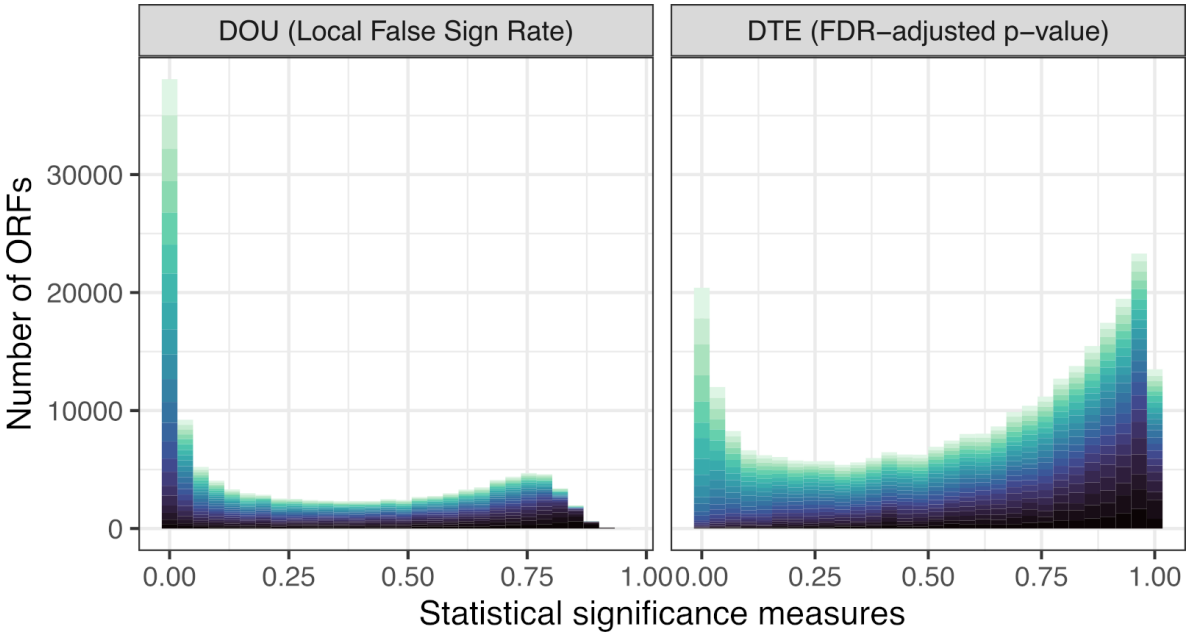

Figure 5: `DOTSeq DOU` module is well-calibrated for differential ORF usage analysis. Distributions of aggregated Local False Sign Rate (LFSR) and FDR-adjusted p-values across 11 simulated scenarios for the `DOU` and `DTE` modules, respectively. These scenarios model DOU with varying absolute log-fold change differences between uORFs and mORFs from 1 to 3. Bar height indicates the total number of ORFs across all scenarios, whereas color gradient indicates the composition of scenarios contributing to the corresponding p-value bin.The `DOU` module shows a right-skewed distribution, consistent with well-calibrated statistical significance. In contrast, the `DTE` module exhibits a U-shaped distribution, indicating p-value inflation.

Table 1: `DOTSeq` statistical modules for Differential ORF Usage (DOU) and Differential Translation Efficiency (DTE).

Feature	`DOU` **module**	`DTE` **module**
Biological question	Does the relative translation of one ORF compared to other ORFs within the same gene change across conditions?	Does the translation efficiency of a specific ORF change across conditions?
Model type	Beta-binomial GLM/GLMM via `glmmTMB`.	Negative binomial GLM via `DESeq2`.
Response variable	Proportion of counts for an ORF within its gene, modelled as a beta-binomial response (y_i, $n_i - y_i$), where y_i is the count for ORF_i (Ribo-seq or RNA-seq), and n_i is the total count for all ORFs in the same gene.	Raw ORF-level counts for Ribo-seq and RNA-seq, modelled directly without normalisation to gene-level totals.
Library size normalisation	No normalisation needed as proportional modelling is robust to variations in library size.	Handled internally by `DESeq2::estimateSizeFactors`.
Dispersion modelling	Library strategy-specific dispersion via dispersion formula, where each ORF has separate dispersion parameters for Ribo-seq and RNA-seq.	Shared dispersion across strategies, where each ORF has a single dispersion parameter. Dispersion is estimated and shrunk based on a mean-dispersion trend.
Interaction term usage	`condition:strategy` in the mean formula. Captures changes in the expected proportion of counts for one ORF relative to other ORFs within the same gene, across both Ribo-seq and RNA-seq data.	`condition:strategy` in the design formula. Captures changes in the ratio of Ribo-seq to RNA-seq counts for individual ORFs.
Hypothesis tests	LRT and Wald test.	Wald test (`DESeq2` default).
Post hoc contrasts	Via `emmeans`	Via `DESeq2::results`.
Empirical Bayesian shrinkage	Adaptive shrinkage for effect sizes (log-odds changes) via `ashr`	Approximate/adaptive shrinkage effect sizes (log2-fold changes) via `apeglm`/`ashr`.